BRAVE🦁: Broadening the visual encoding of vision-language models

Oğuzhan Fatih Kar   Alessio Tonioni   Petra Poklukar

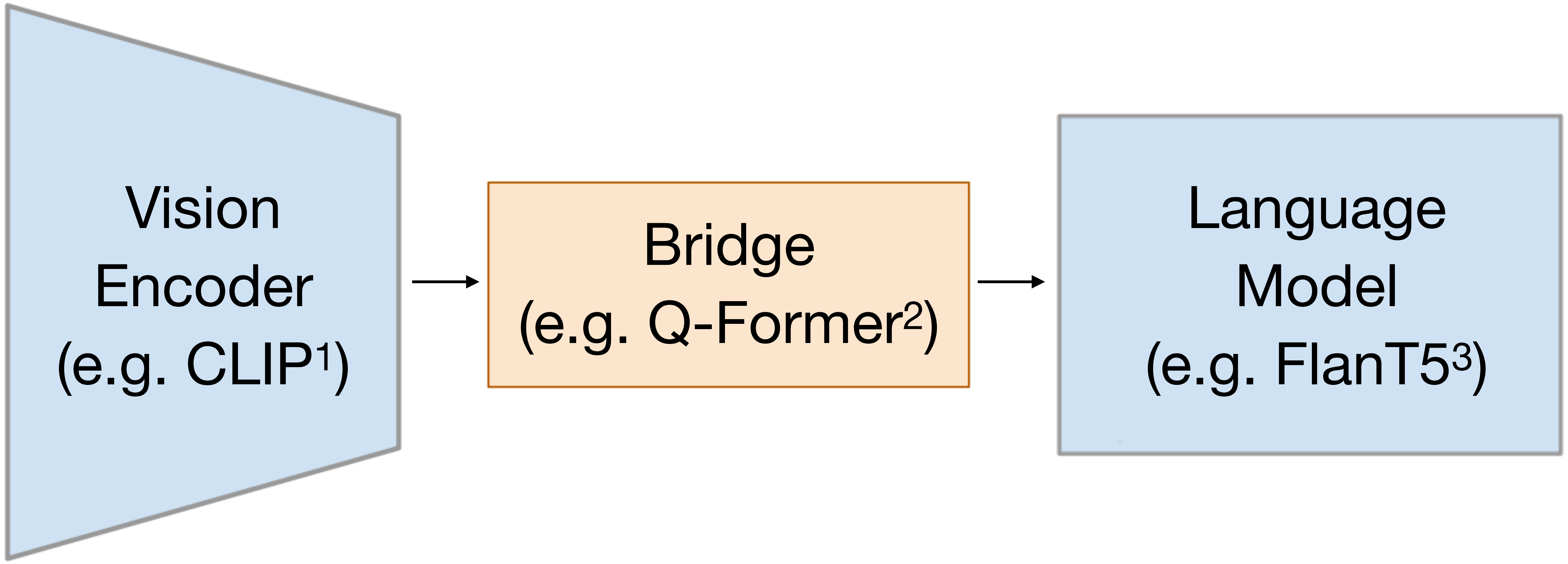
Achin Kulshrestha   Amir Zamir   Federico Tombari

ECCV 2024 (Oral)

brave-vlms.epfl.ch

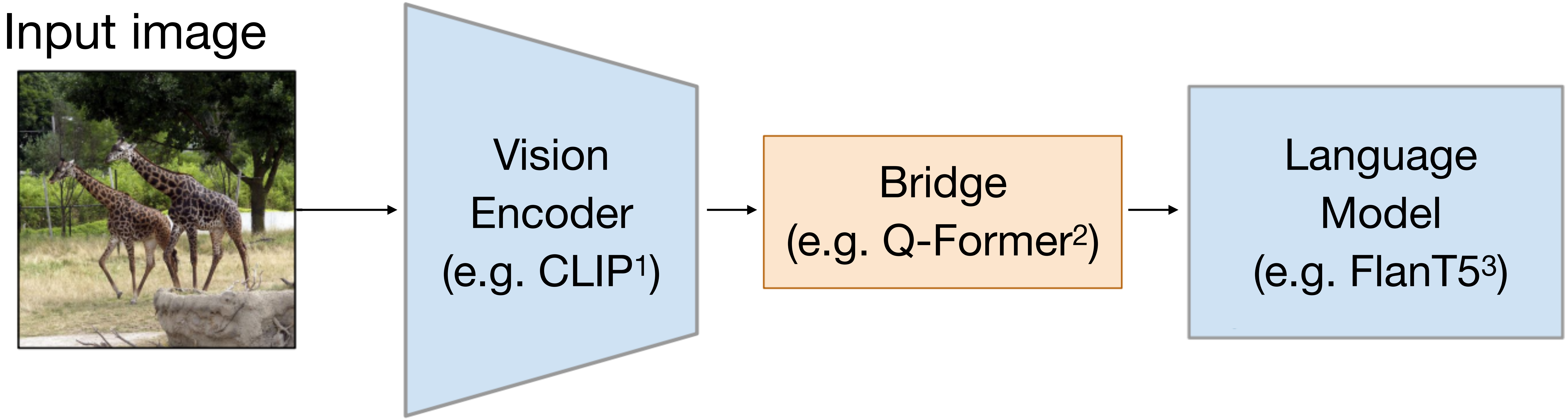
Swiss Federal
Institute of
Technology

Vision-Language Models (VLMs)


Vision Encoder (e.g. CLIP[1])
Bridge (e.g. Q-Former[2])
Language Model (e.g. FlanT5[3])

[1] *Radford et al. 2021*
[2] *Li et al. 2023*
[3] *Chung et al. 2023*

Vision-Language Models (VLMs)



Input image

Vision Encoder (e.g. CLIP[1])

Bridge (e.g. Q-Former[2])

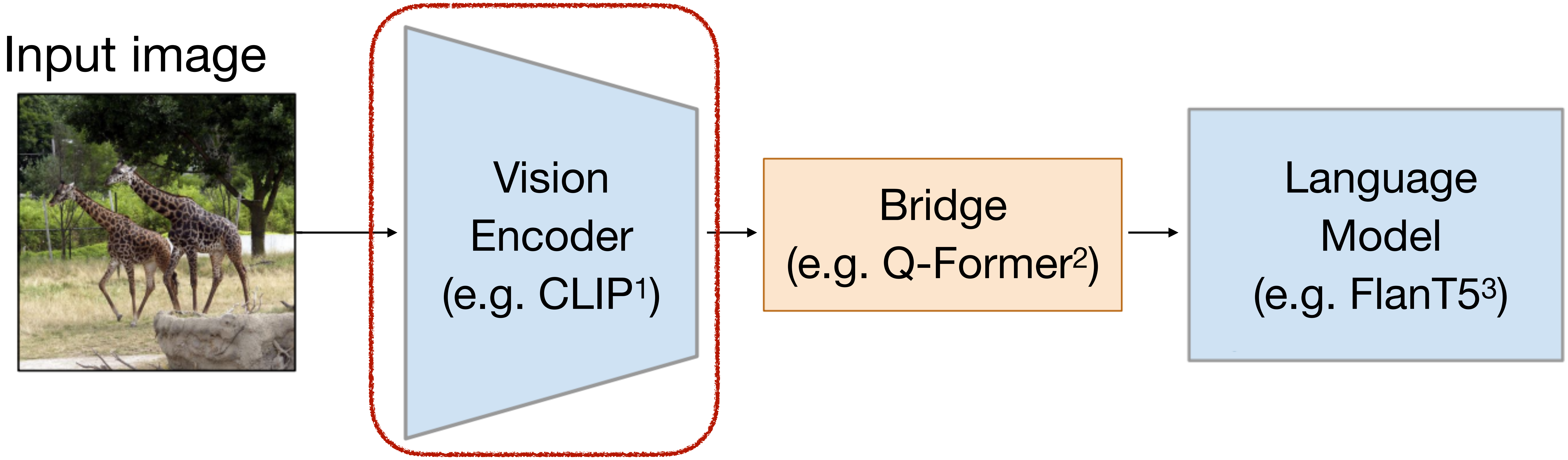
Language Model (e.g. FlanT5[3])

[1] Radford et al. 2021

[2] Li et al. 2023

[3] Chung et al. 2023

Vision-Language Models (VLMs)



Input image

Vision Encoder (e.g. CLIP[1])

Bridge (e.g. Q-Former[2])

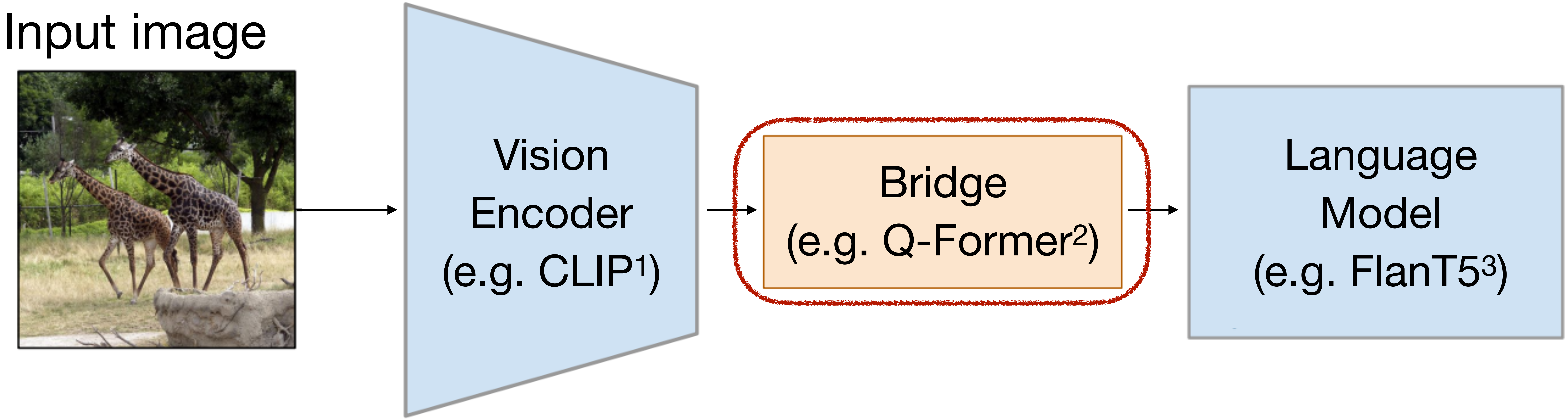
Language Model (e.g. FlanT5[3])

[1] *Radford et al. 2021*

[2] *Li et al. 2023*

[3] *Chung et al. 2023*

Vision-Language Models (VLMs)



Input image

Vision Encoder (e.g. CLIP[1])

Bridge (e.g. Q-Former[2])

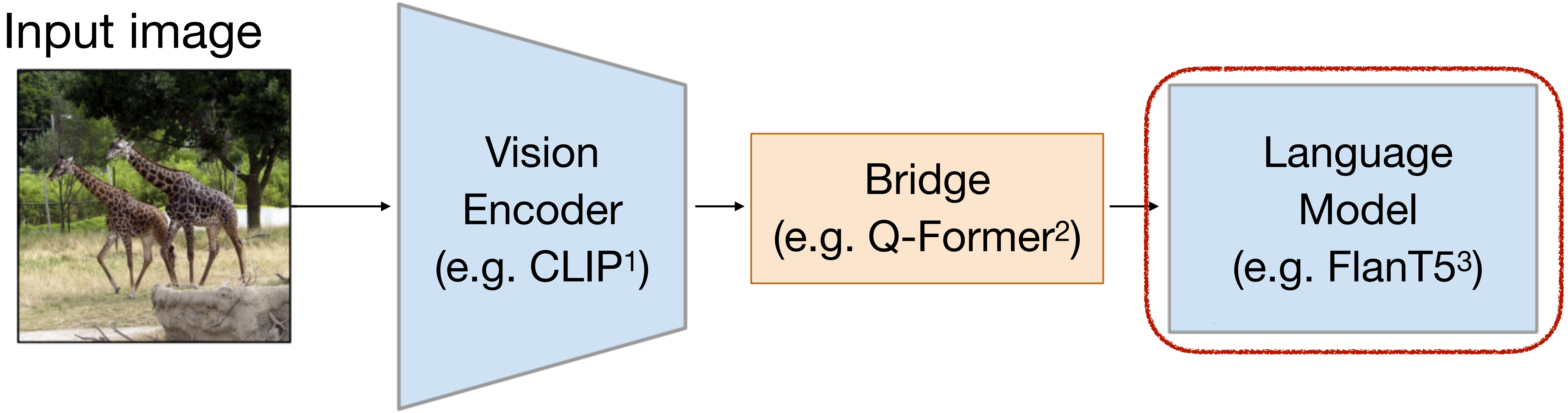
Language Model (e.g. FlanT5[3])

[1] *Radford et al. 2021*

[2] *Li et al. 2023*

[3] *Chung et al. 2023*

Vision-Language Models (VLMs)

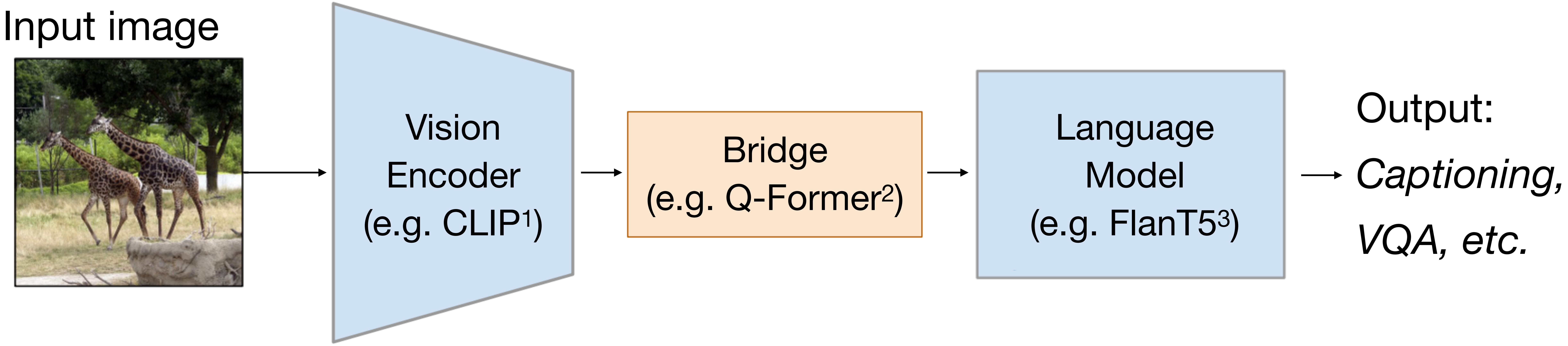

Input image
Vision Encoder (e.g. CLIP[1])
Bridge (e.g. Q-Former[2])
Language Model (e.g. FlanT5[3])

[1] *Radford et al. 2021*

[2] *Li et al. 2023*

[3] *Chung et al. 2023*

Vision-Language Models (VLMs)



Input image

Vision Encoder (e.g. CLIP[1])

Bridge (e.g. Q-Former[2])

Language Model (e.g. FlanT5[3])

Output: *Captioning, VQA, etc.*

[1] *Radford et al. 2021*

[2] *Li et al. 2023*

[3] *Chung et al. 2023*

VLMs have important shortcomings

VLMs have important shortcomings

- Limited language capabilities
  - Hallucinations[1,2]
  - Logical faults[3,4]

[1] *Bang et al. 2023*
[2] *Guo et al. 2023*
[3] *Shen et al. 2023*
[4] *Thorp et al. 2023*

VLMs have important shortcomings

- Limited language capabilities
  - Hallucinations[1,2]
  - Logical faults[3,4]

- Limited visual understanding
  - “Blindness”[5]
  - Visual hallucinations[6]

[1] Bang et al. 2023
[2] Guo et al. 2023
[3] Shen et al. 2023
[4] Thorp et al. 2023
[5] Tong et al. 2024
[6] Li et al. 2023

VLMs have important shortcomings

- Limited language capabilities
  - Hallucinations[1,2]
  - Logical faults[3,4]

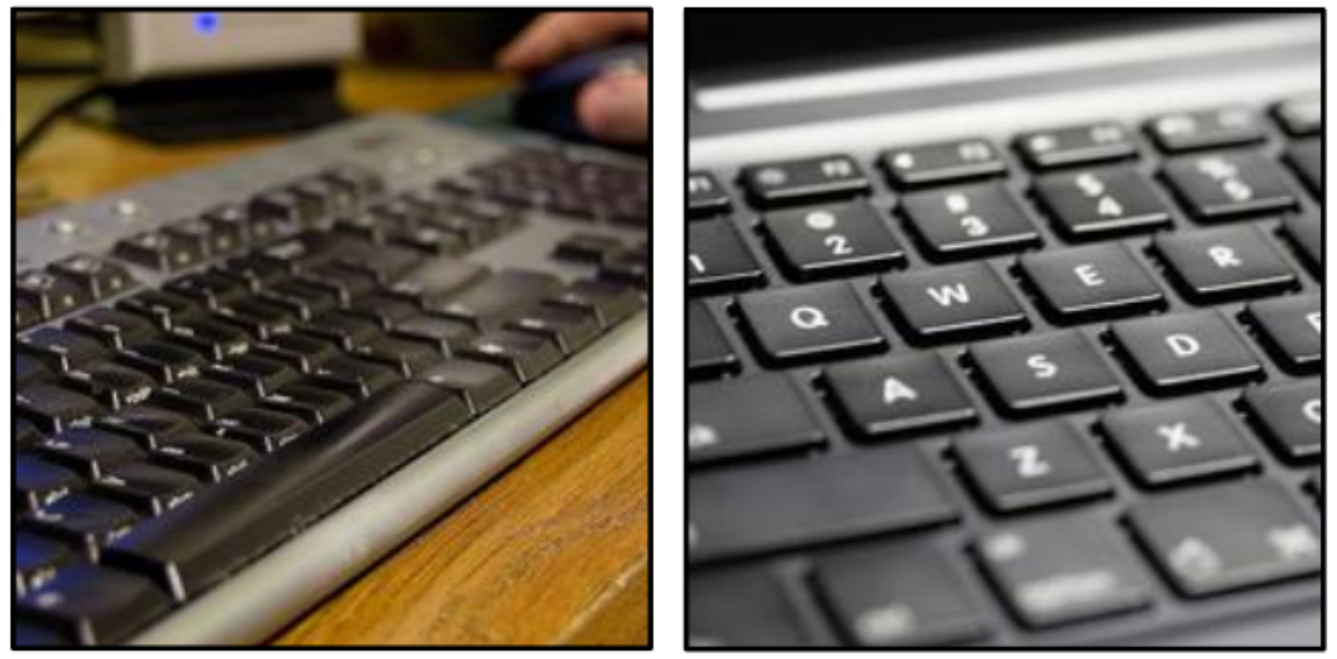
- Limited visual understanding
  - “Blindness”[5]
  - Visual hallucinations[6]

← **Our focus**

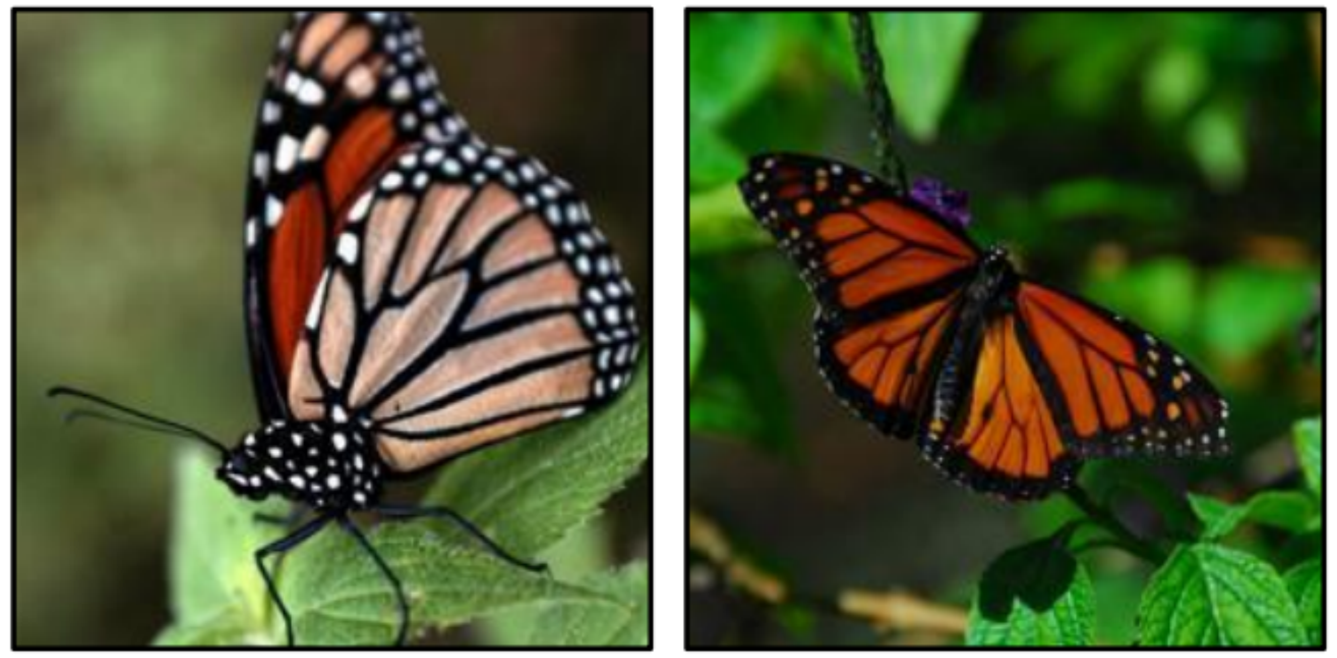
[1] Bang et al. 2023
[2] Guo et al. 2023
[3] Shen et al. 2023
[4] Thorp et al. 2023
[5] Tong et al. 2024
[6] Li et al. 2023

Example failures — confusing image pairs[1]

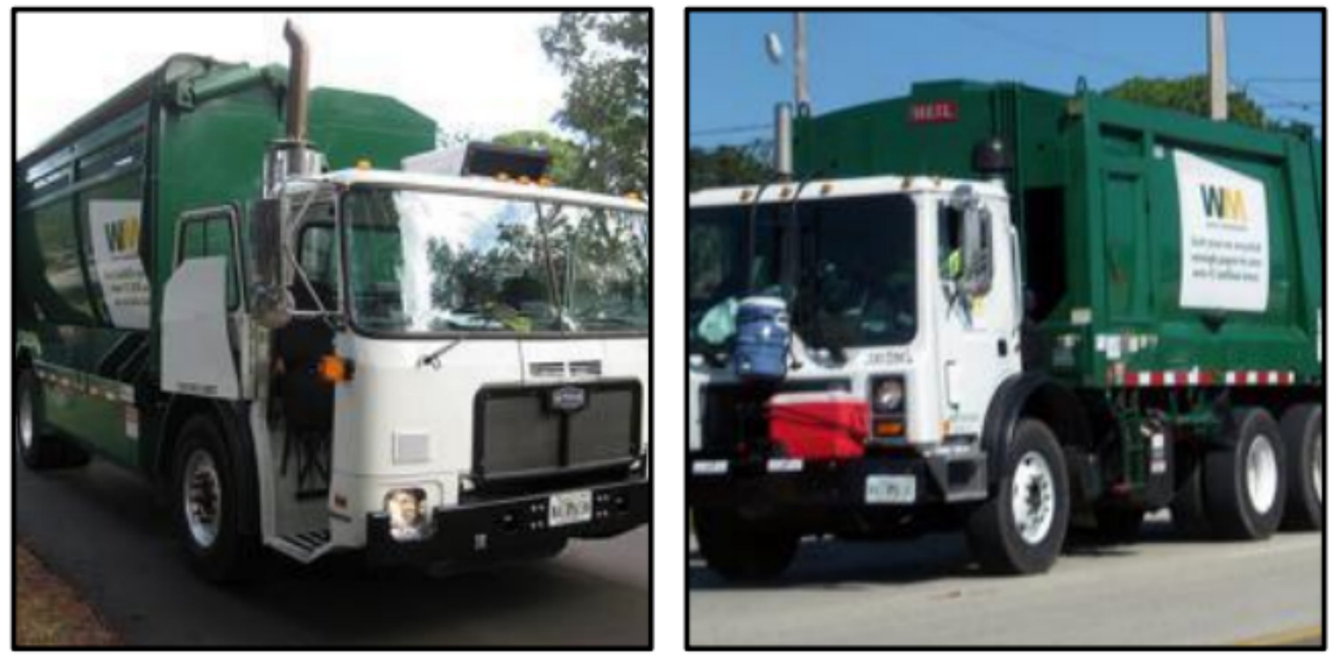
Is there a hand using the mouse in this image?



Are the butterfly's feet visible?



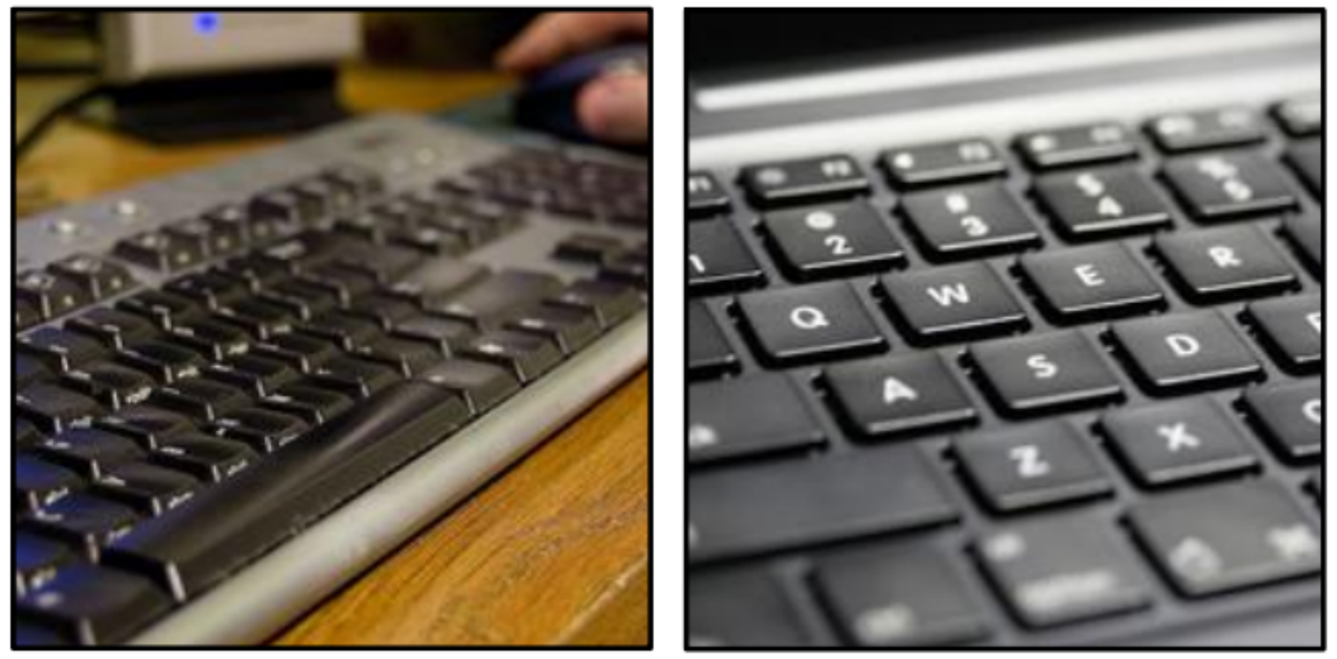
Is the door of the truck open?



[1] Tong et al. 2024

Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?



Are the butterfly's feet visible?

Is the door of the truck open?

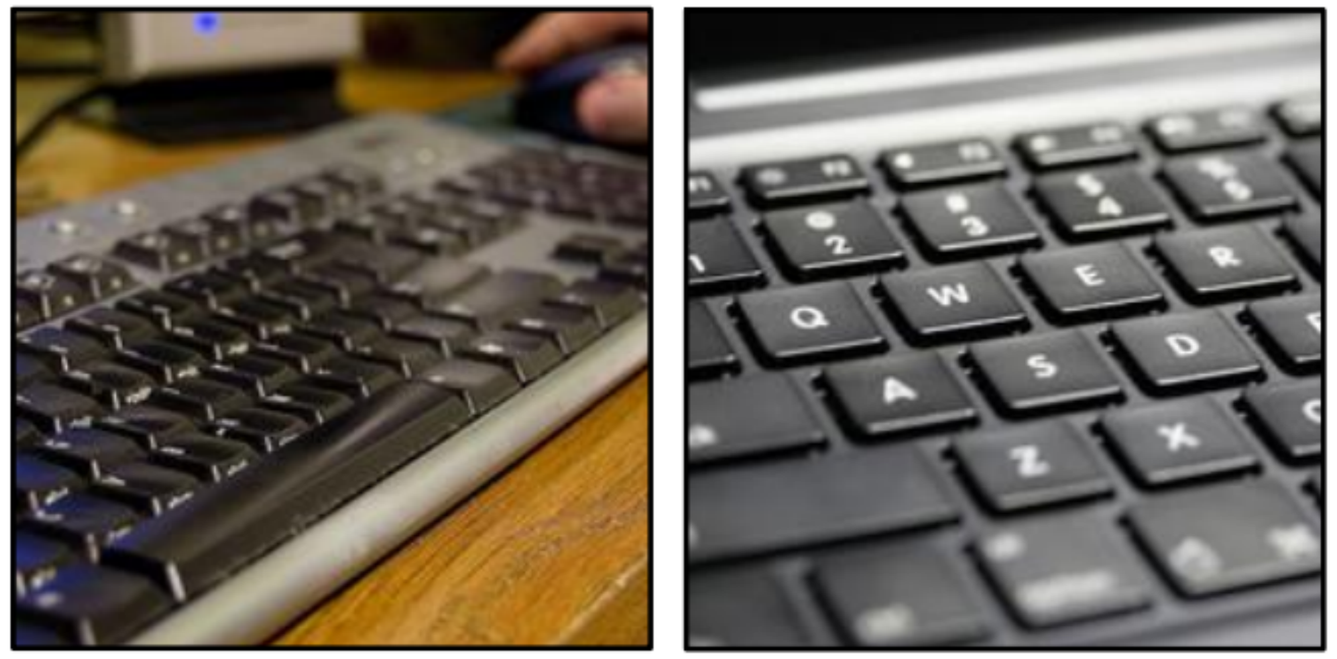
InstructBLIP :

LLaVA-1.5 :

[1] *Tong et al. 2024*

Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?



Yes *No*

Are the butterfly's feet visible?

Is the door of the truck open?

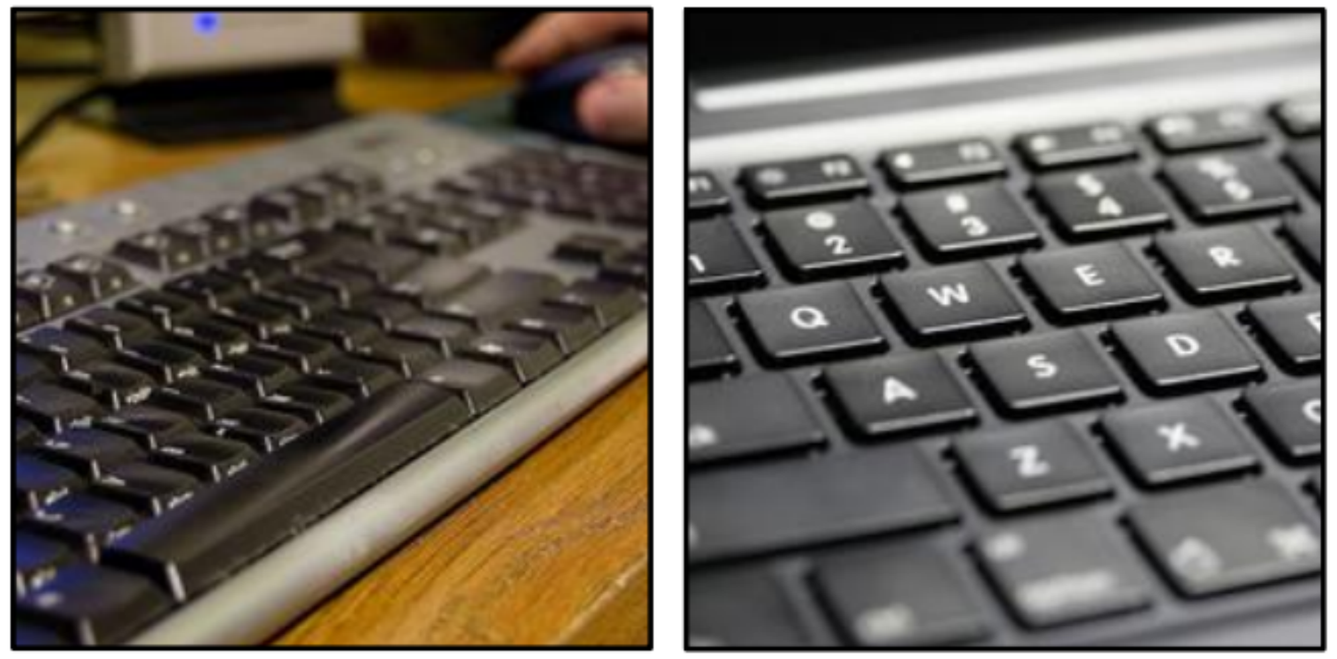
InstructBLIP :

LLaVA-1.5 :

[1] *Tong et al. 2024*

Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?



Yes *No*

InstructBLIP :	Yes	No	✓
LLaVA-1.5 :	No	No	✗

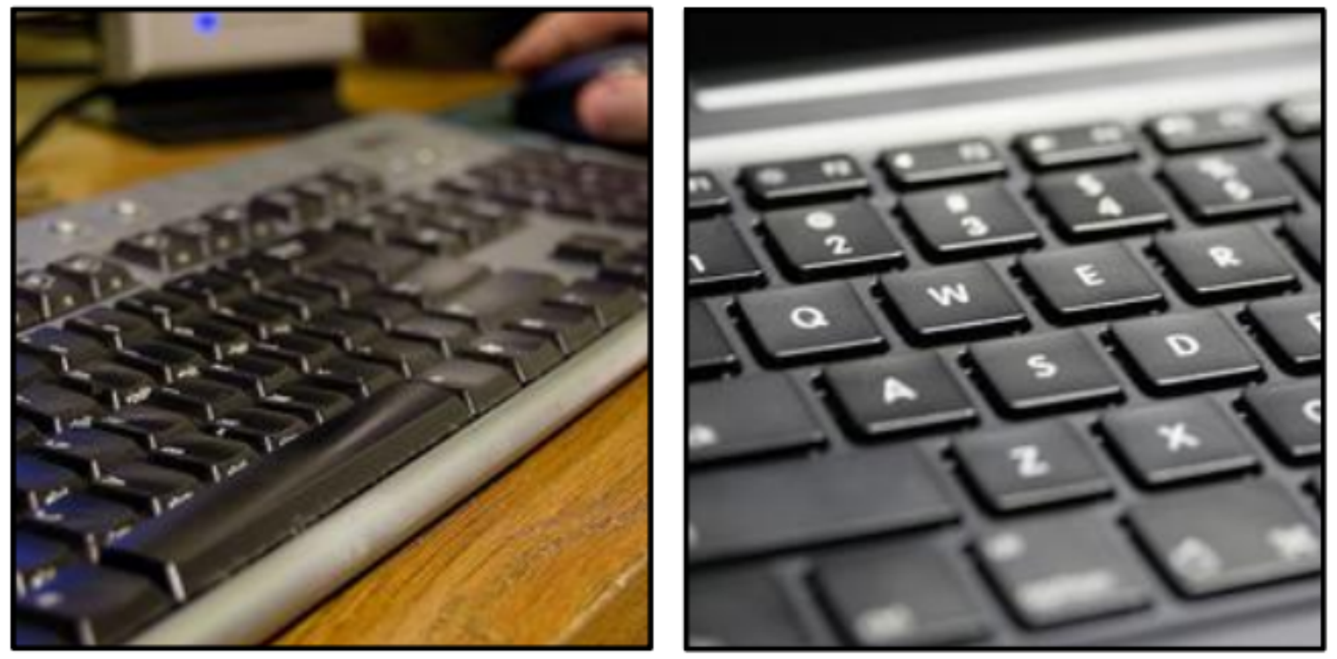
Are the butterfly's feet visible?

Is the door of the truck open?

[1] *Tong et al. 2024*

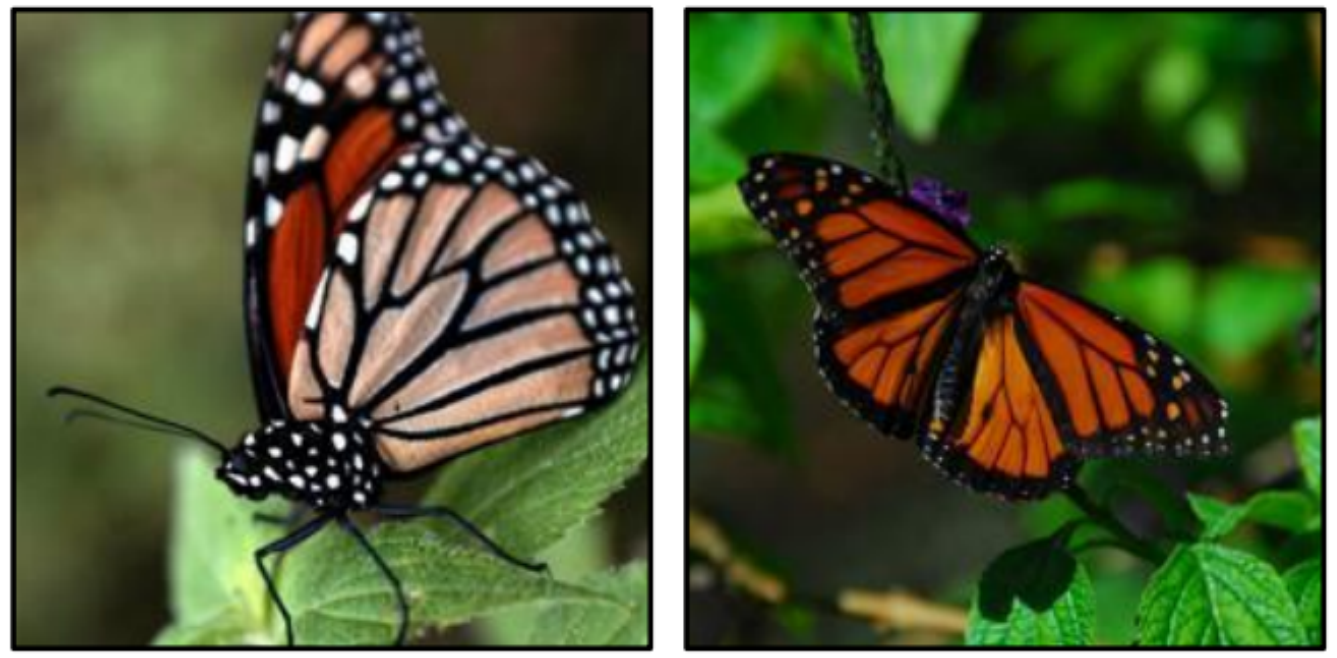
Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?



Yes *No*

Are the butterfly's feet visible?



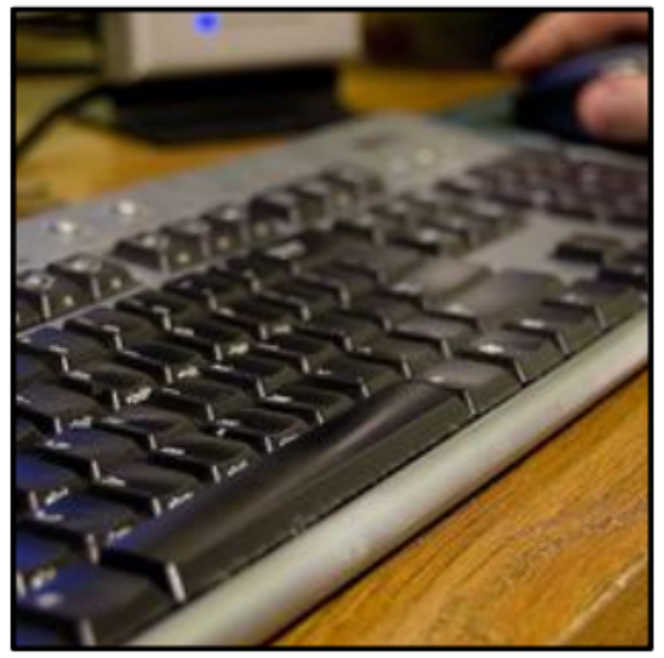
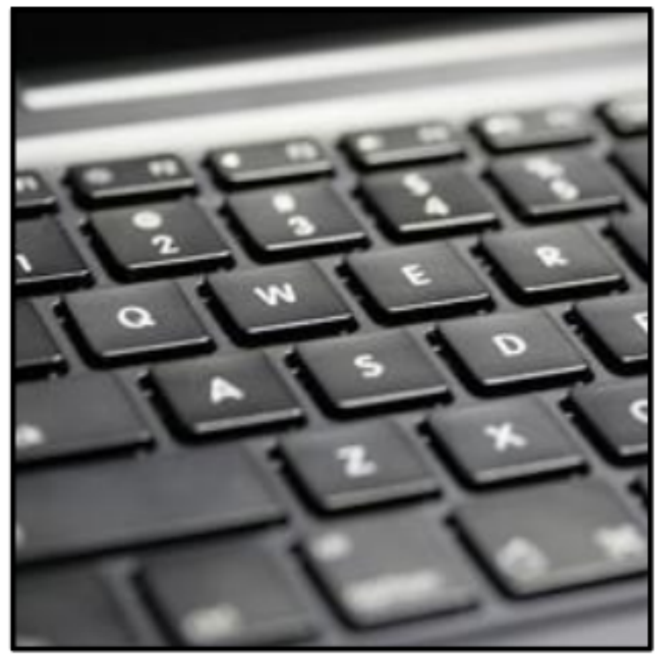
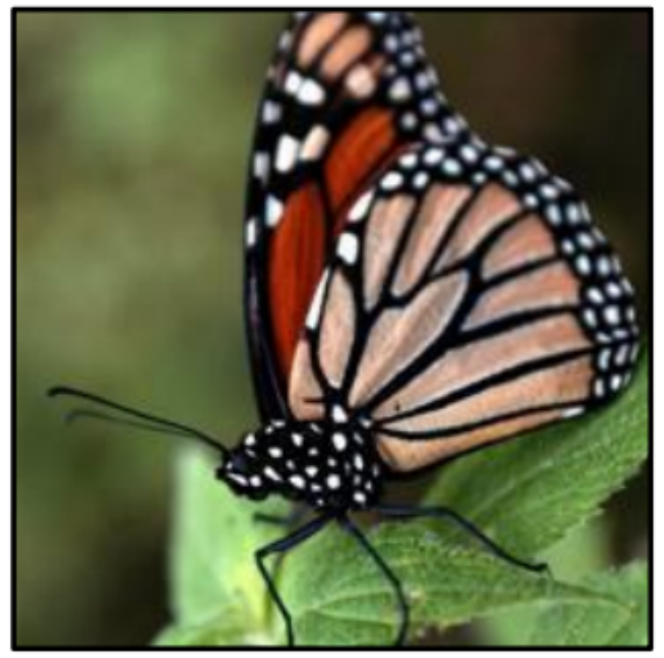
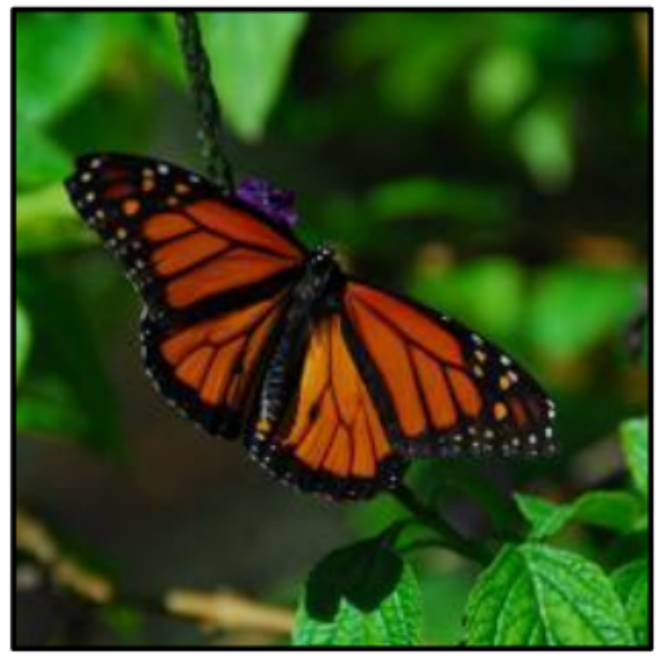
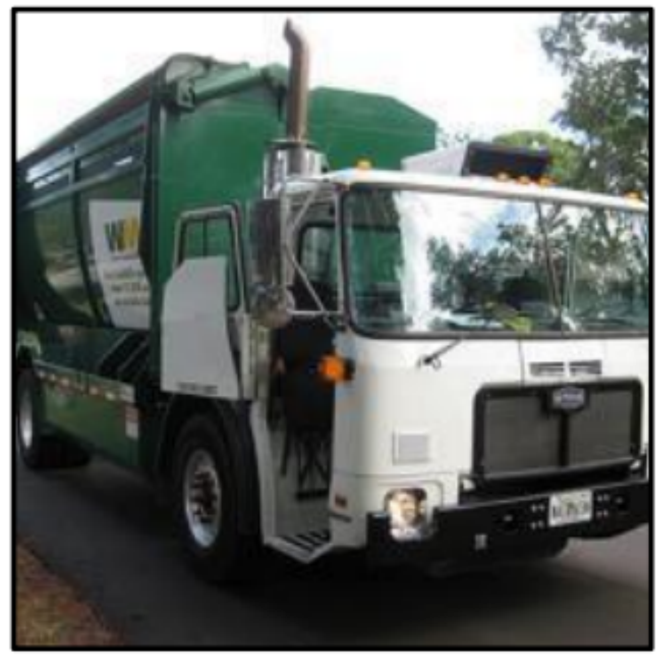
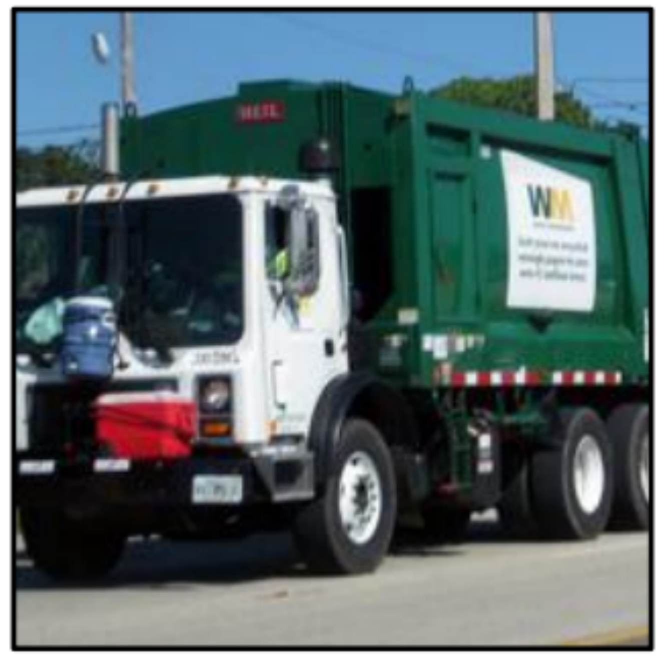
Yes *No*

Is the door of the truck open?

InstructBLIP :	Yes	No ✅	Yes	Yes ❌
LLaVA-1.5 :	No	No ❌	Yes	No ✅

[1] *Tong et al. 2024*

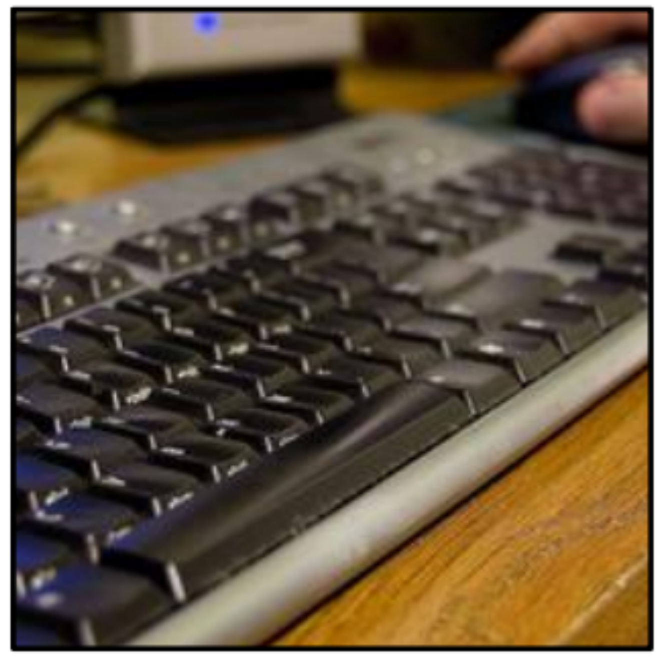
Example failures — confusing image pairs[1]

	Is there a hand using the mouse in this image?		*Are the butterfly's feet visible?*		*Is the door of the truck open?*	
						
	Yes	*No*	*Yes*	*No*	*Yes*	*No*
InstructBLIP :	Yes	No ✅	Yes	Yes ❌	Yes	Yes ❌
LLaVA-1.5 :	No	No ❌	Yes	No ✅	Yes	Yes ❌

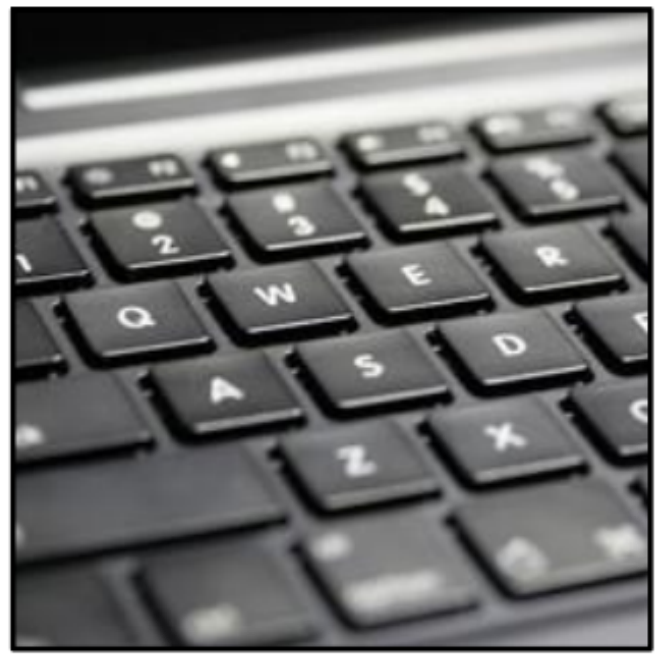
[1] *Tong et al. 2024*

Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?

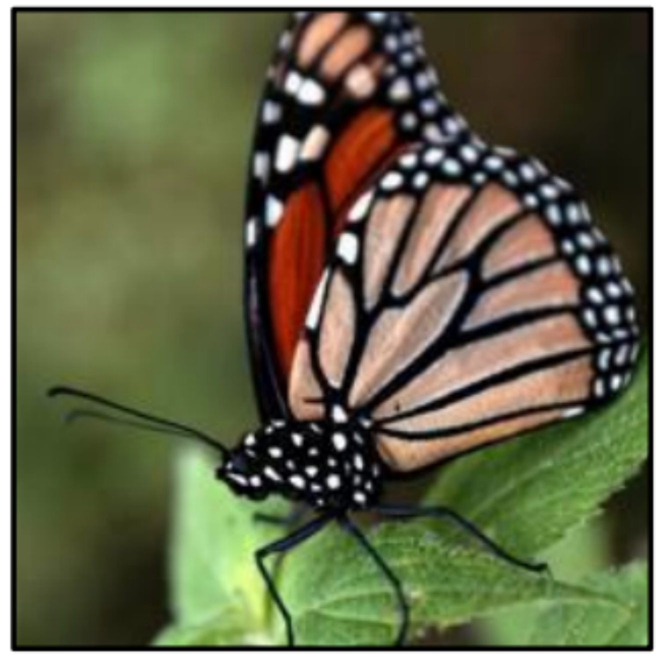


Yes

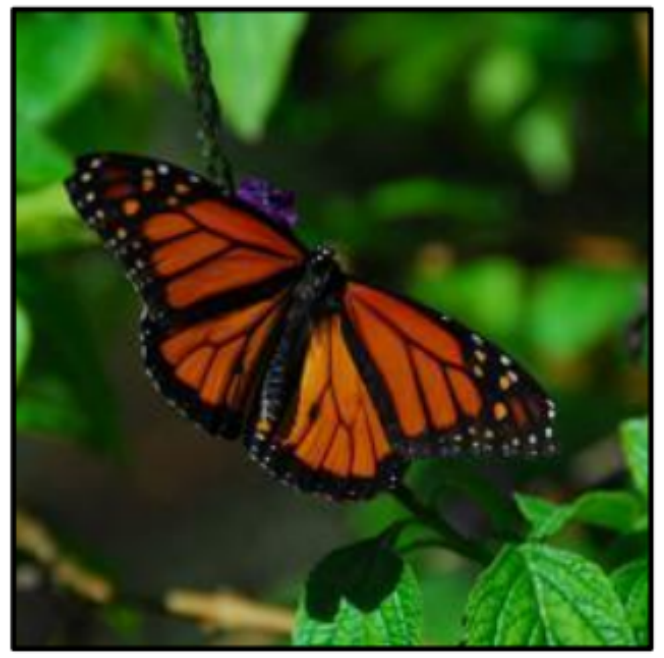


No

Are the butterfly's feet visible?

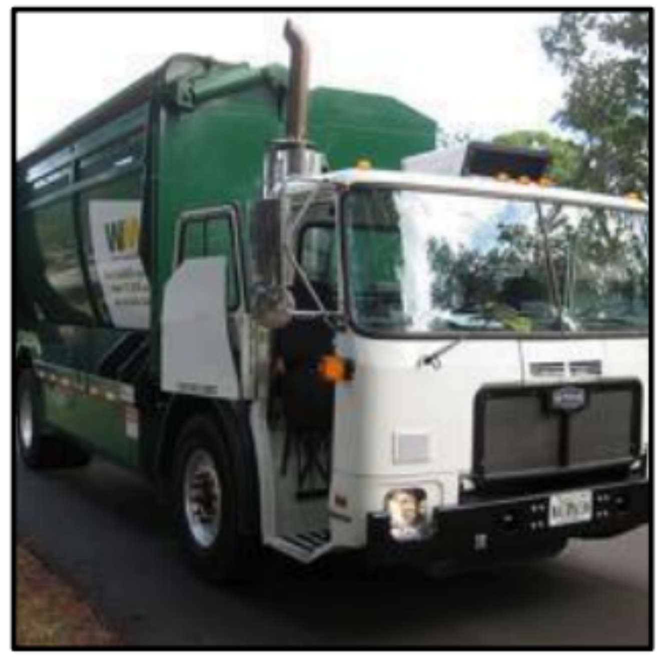


Yes

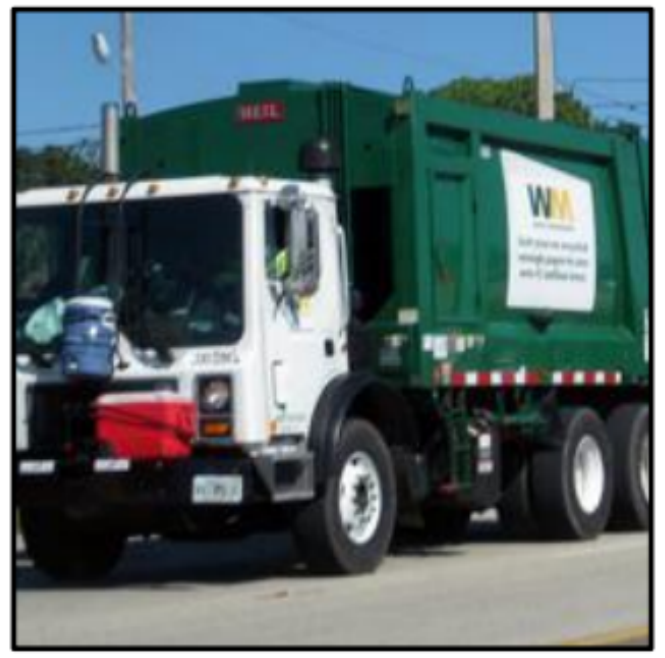


No

Is the door of the truck open?



Yes



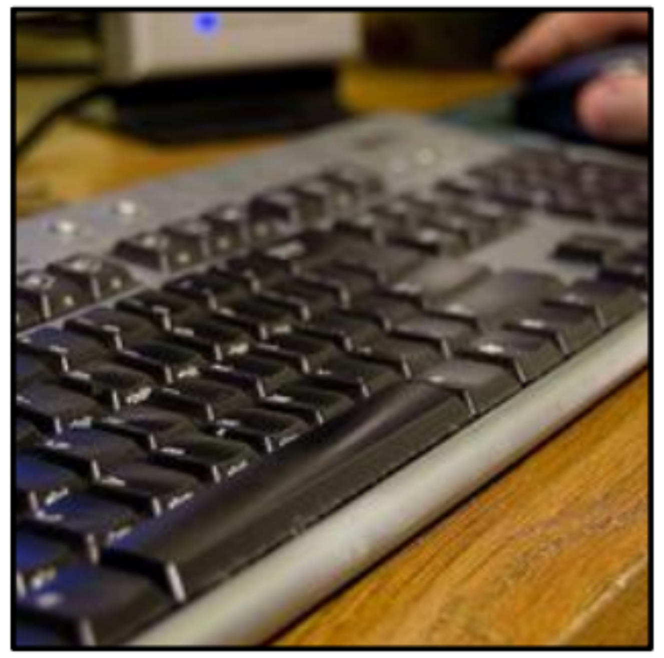
No

	Yes	*No*	*Yes*	*No*	*Yes*	*No*
InstructBLIP : (EVA Encoder)	Yes	No ✅	Yes	Yes ❌	Yes	Yes ❌
LLaVA-1.5 : (CLIP Encoder)	No	No ❌	Yes	No ✅	Yes	Yes ❌

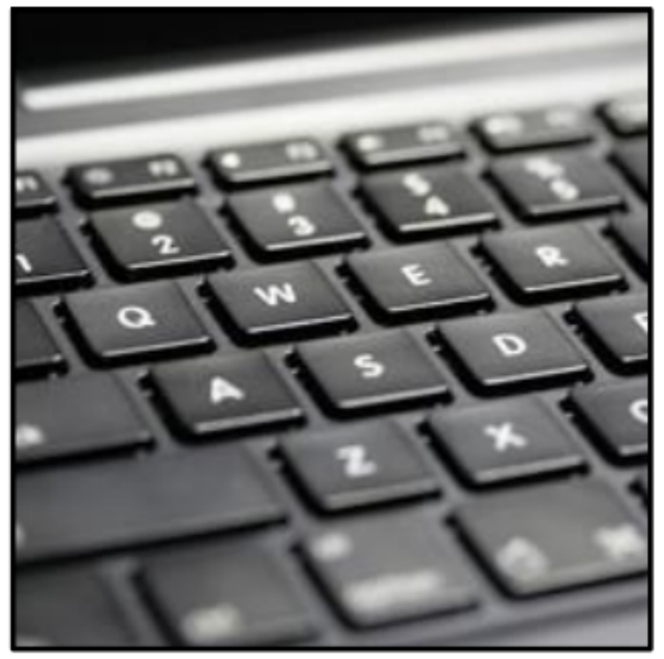
[1] *Tong et al. 2024*

Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?

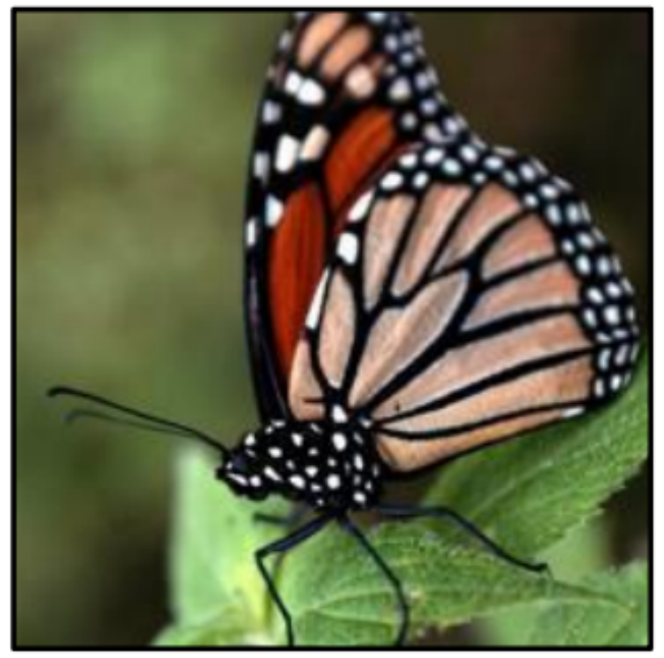


Yes

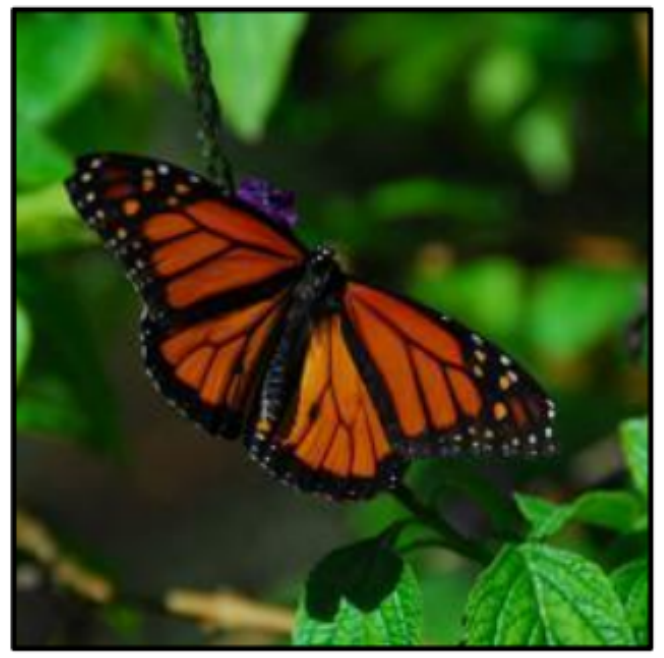


No

Are the butterfly's feet visible?

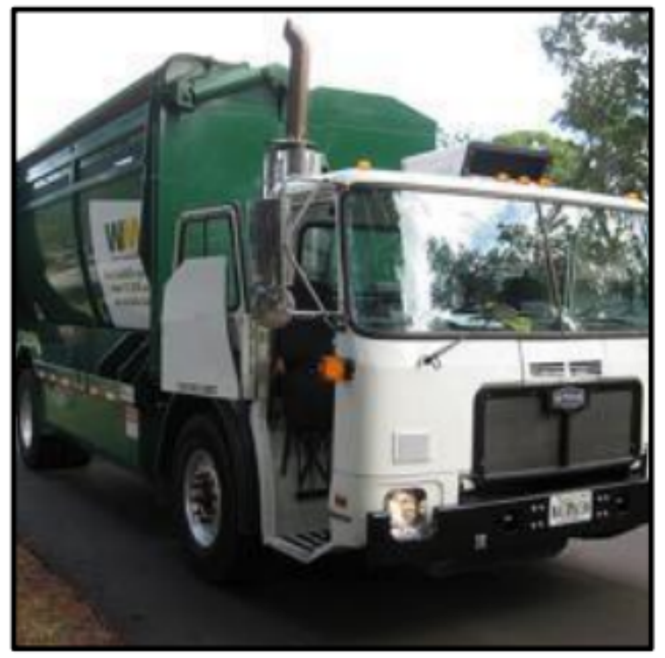


Yes

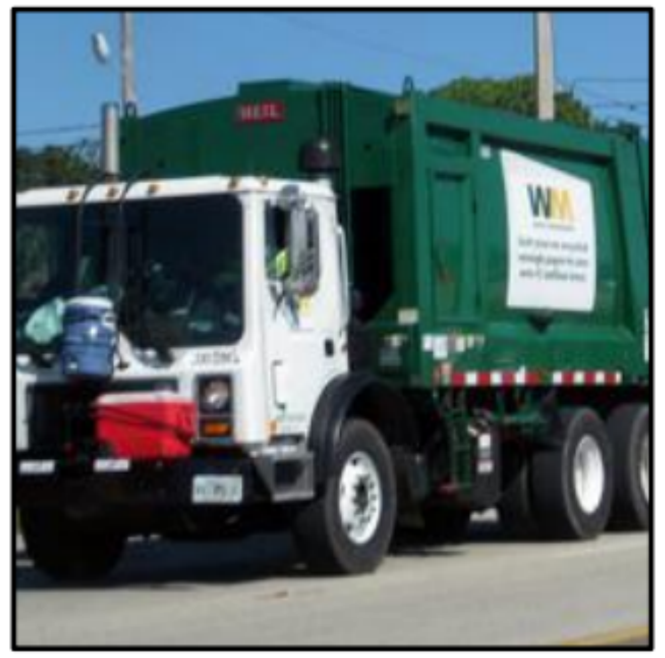


No

Is the door of the truck open?



Yes



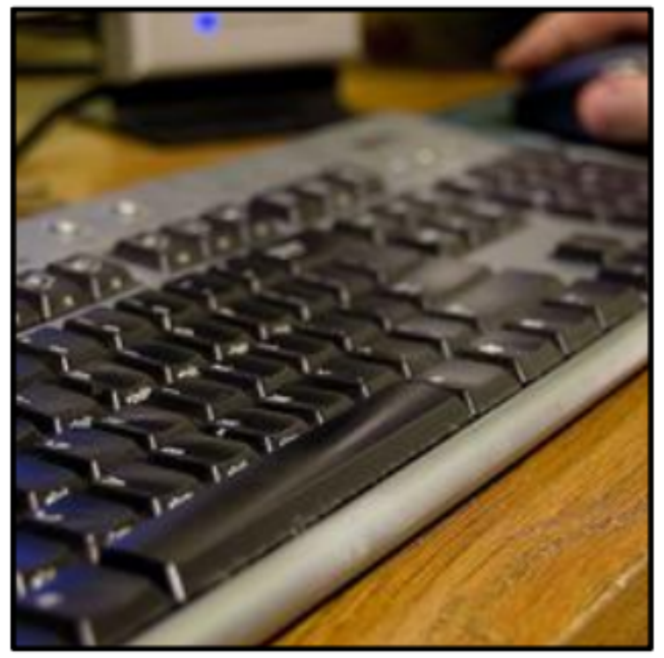
No

InstructBLIP : (Ovr. Acc. 16.7%)	Yes	No ✅	Yes	Yes ❌	Yes	Yes ❌
LLaVA-1.5 : (Ovr. Acc. 24.7%)	No	No ❌	Yes	No ✅	Yes	Yes ❌

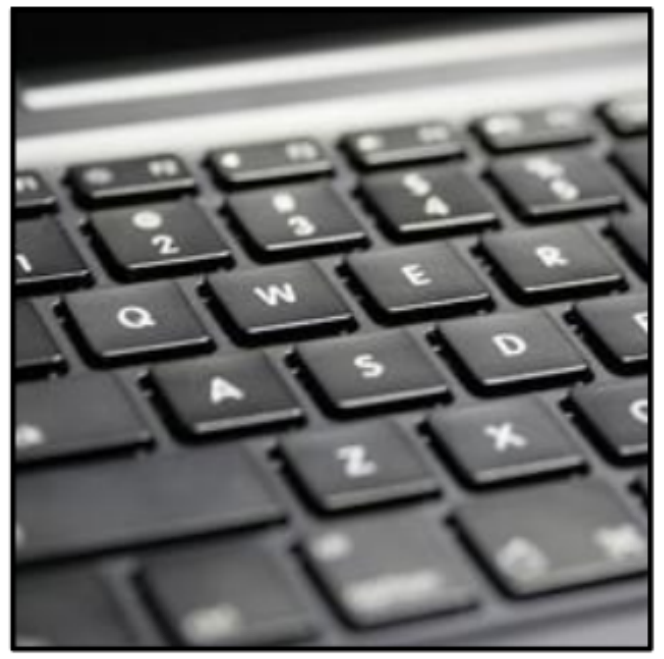
[1] *Tong et al. 2024*

Example failures — confusing image pairs[1]

Is there a hand using the mouse in this image?

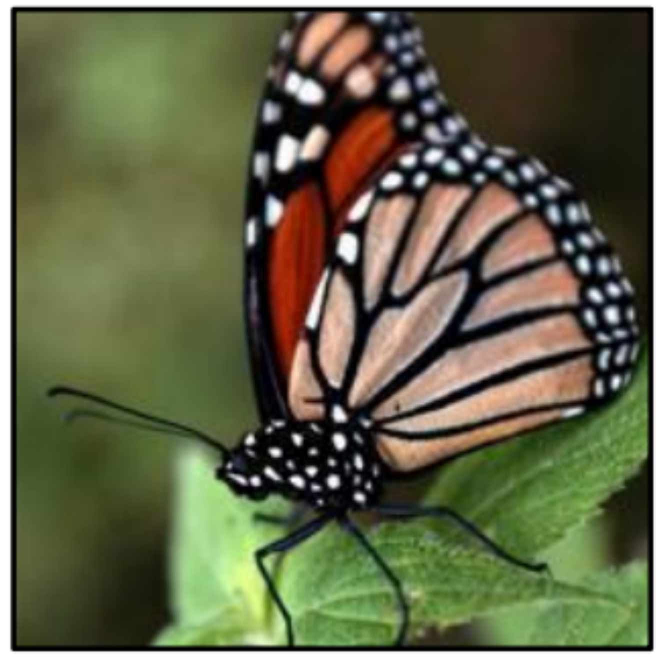


Yes

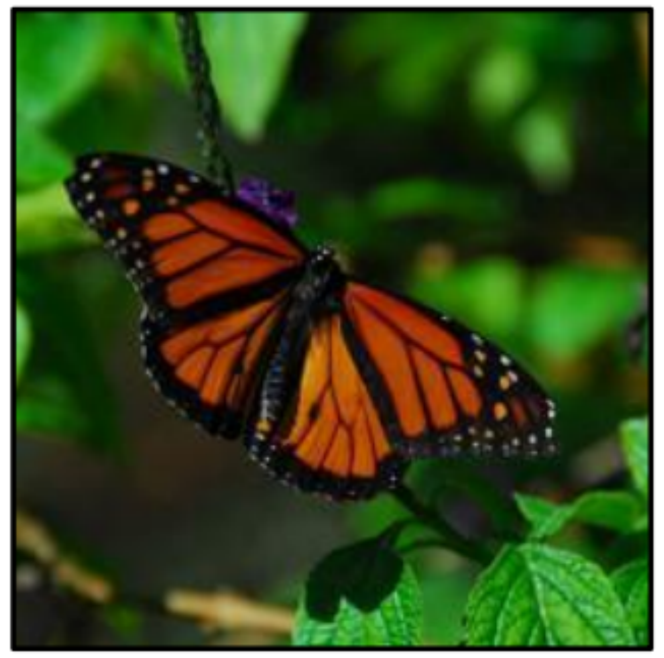


No

Are the butterfly's feet visible?

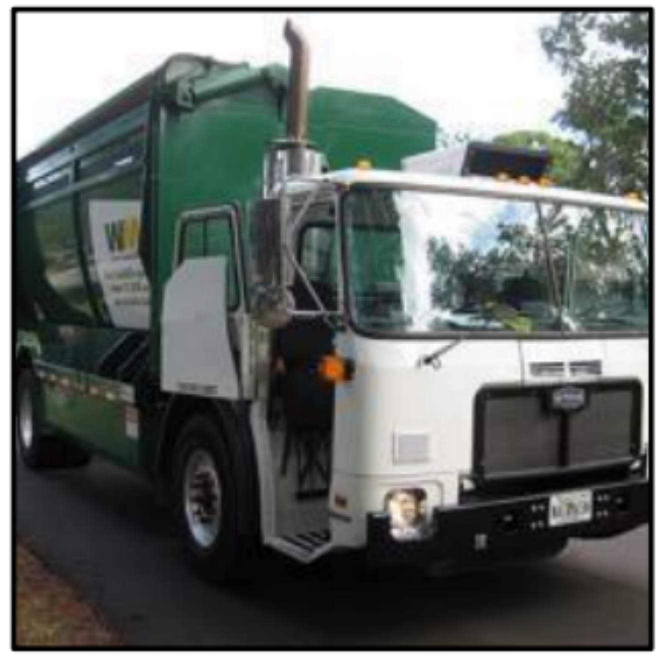


Yes

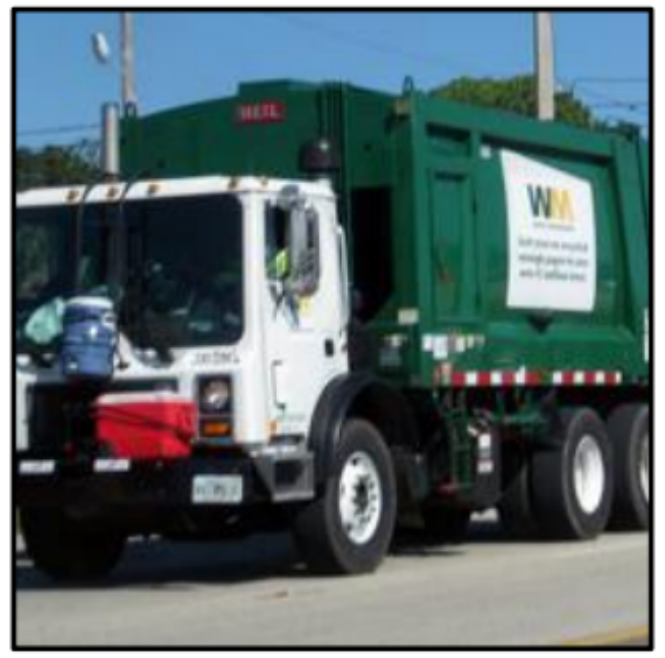


No

Is the door of the truck open?



Yes



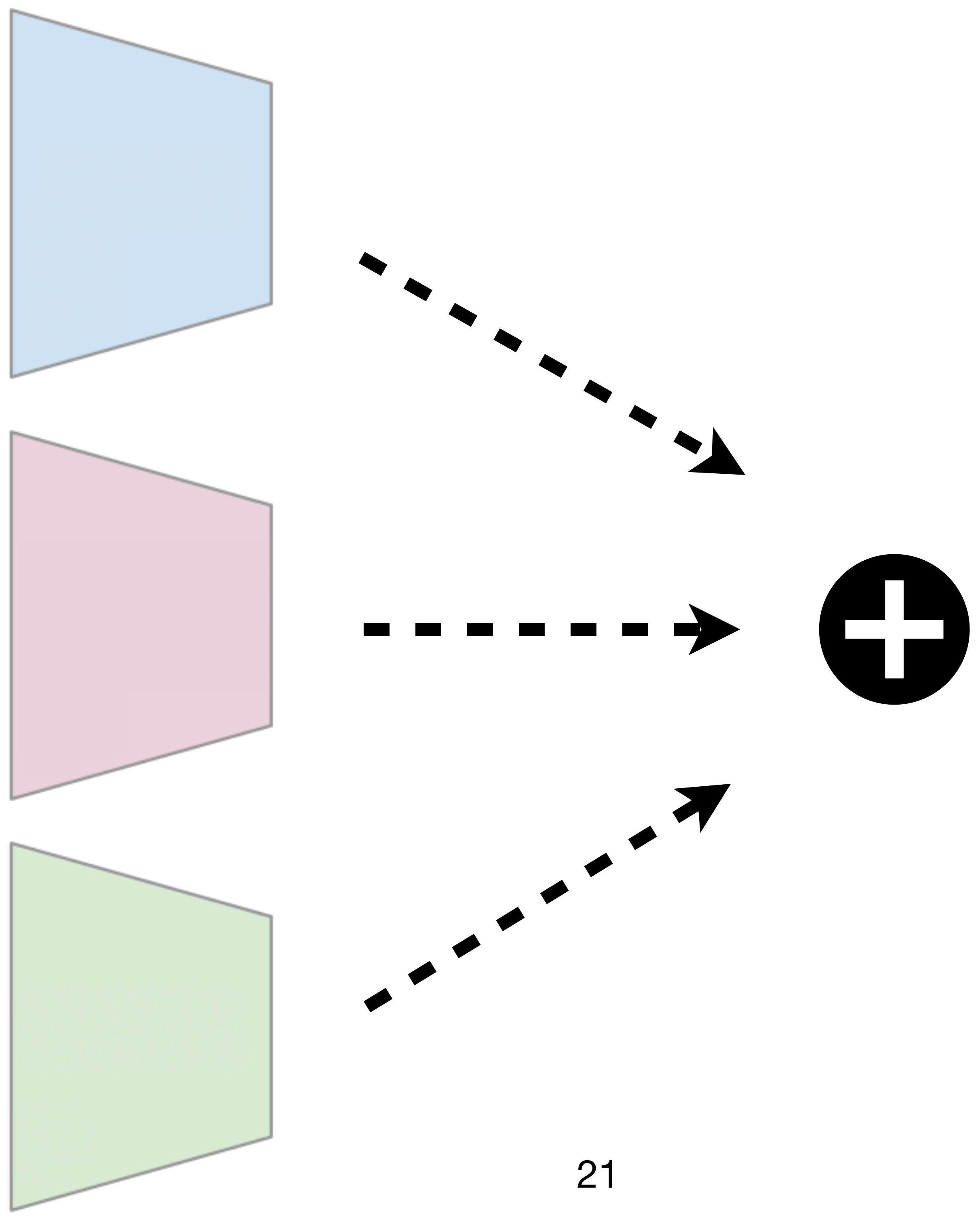
No

InstructBLIP : (Ovr. Acc. 16.7%)	Yes	No ✅	Yes	Yes ❌	Yes	Yes ❌
LLaVA-1.5 : (Ovr. Acc. 24.7%)	No	No ❌	Yes	No ✅	Yes	Yes ❌
BRAVE🦁 : (Ovr. Acc. 42.0%)	Yes	No ✅	Yes	No ✅	Yes	No ✅

[1] *Tong et al. 2024*

BRAVE: Broadening the visual encoding of VLMs

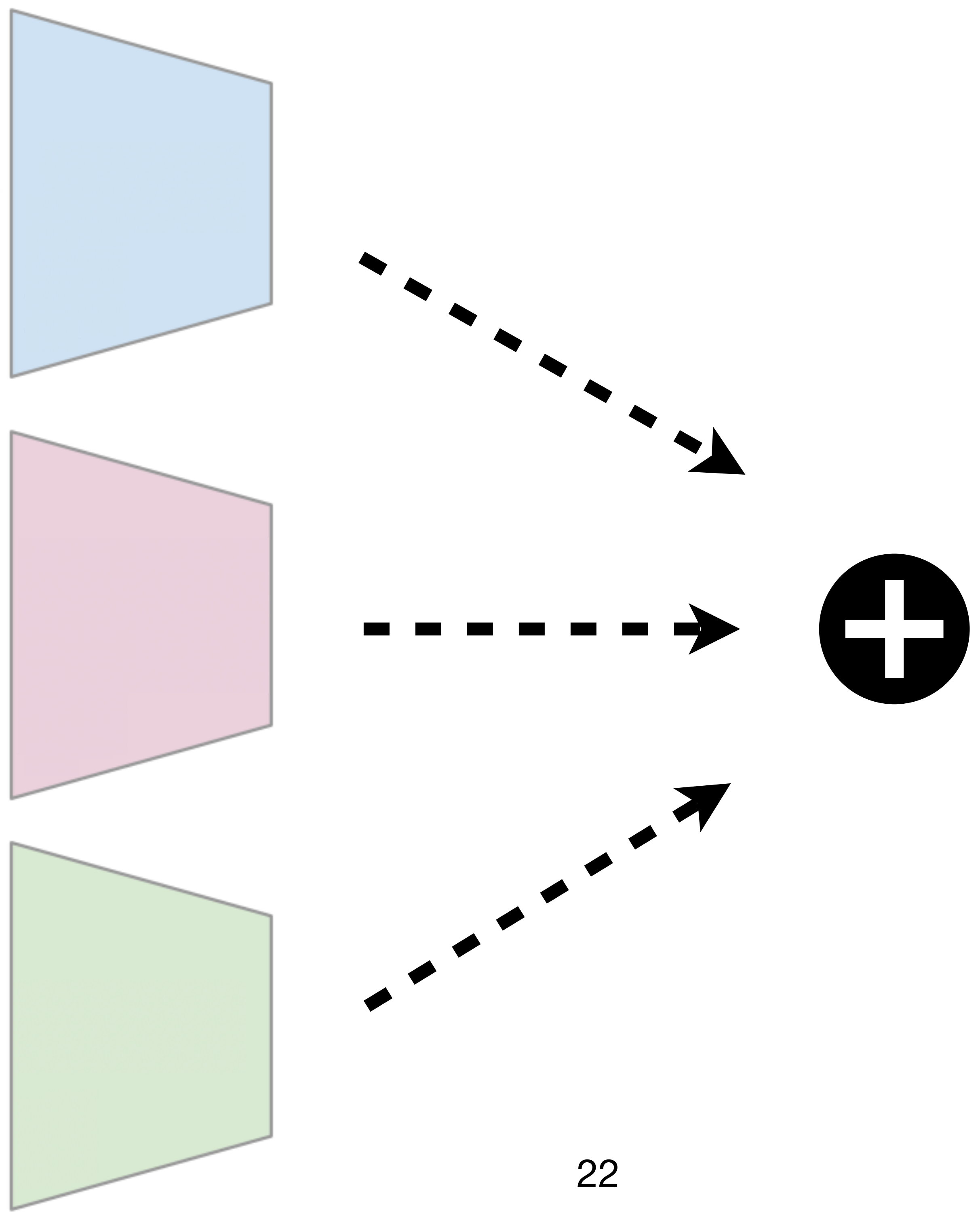
- Core idea from machine learning[1]
  - Different representations -> Different generalization properties
  - Ensemble to create a more complete representation



[1] *Geman et al. 1992*

BRAVE: Broadening the visual encoding of VLMs

- Core idea from machine learning[1]
  - Different representations -> Different generalization properties
  - Ensemble to create a more complete representation
  - Find the strongest set via **benchmarking**



[1] *Geman et al. 1992*

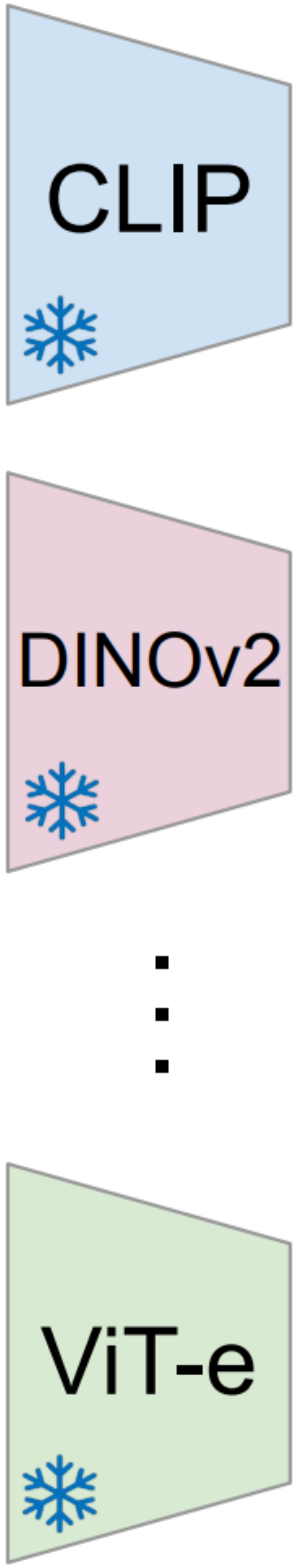
Benchmarking vision encoders

- **8 different encoders**
  - Different objectives
    - Masked modeling, contrastive learning, etc.
  - Different training datasets
    - LAION-2B, JFT-3B, etc.
  - Different model sizes
    - 300M to 4B

Benchmarking vision encoders

- **8 different encoders**
    - Different objectives
        - Masked modeling, contrastive learning, etc.
    - Different training datasets
        - LAION-2B, JFT-3B, etc.
    - Different model sizes
        - 300M to 4B

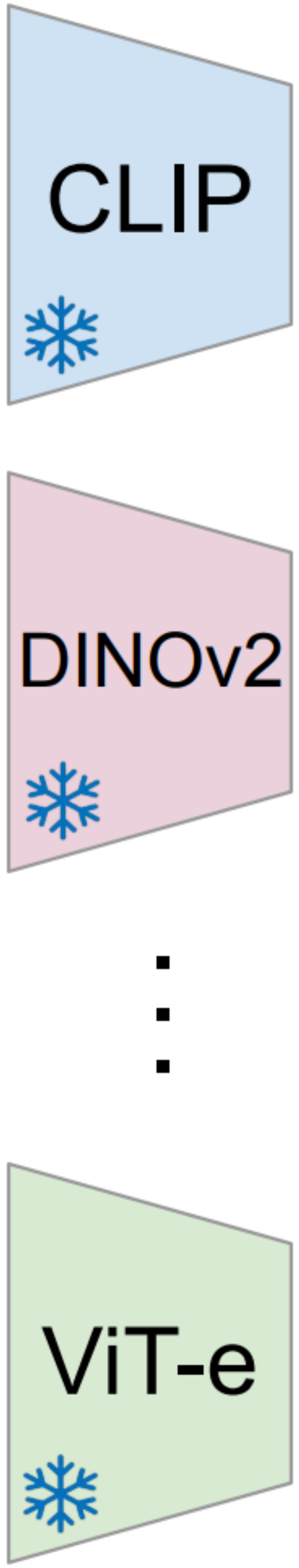
- ***CLIP[1], EVA[2], DINOv2[3], SIGLIP[4], OpenCLIP[5], SILC[6], ViT-e[7], ViT-G[8]***
- **Evaluation tasks: Captioning, VQA**



[1] Radford et al. 2021
[2] Fang et al. 2023
[3] Oquab et al. 2023
[4] Zhai et al. 2023
[5] Cherti et al. 2023
[6] Naeem et al. 2023
[7] Chen et al. 2022
[8] Zhai et al. 2022

Benchmarking vision encoders
Observations

- **No encoder perform consistently well**

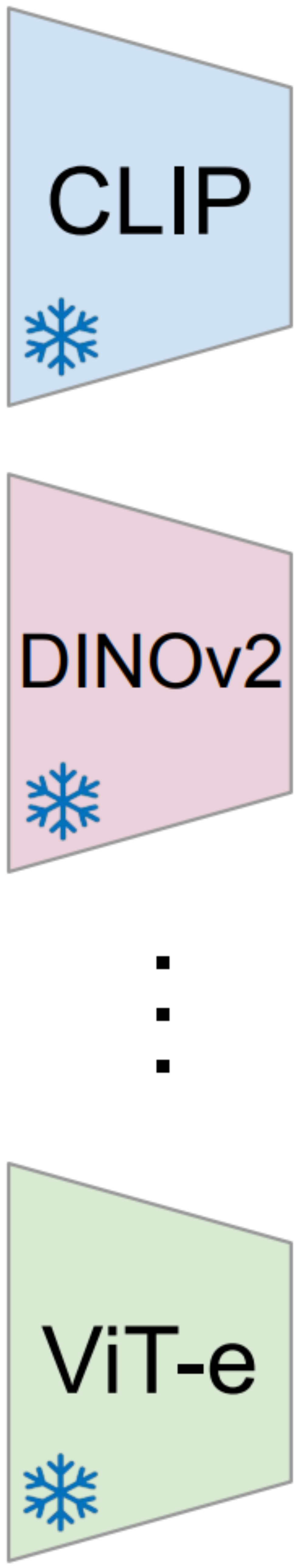


Please see the paper for details

Benchmarking vision encoders
Observations

- **No encoder perform consistently well**
    - Using a single encoder is inherently limited

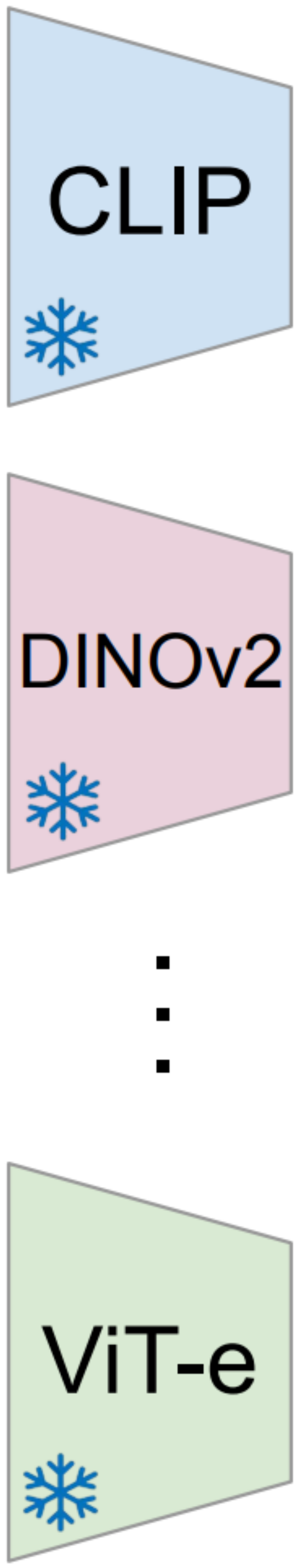
Please see the paper for details



Benchmarking vision encoders
Observations

- **No encoder perform consistently well**
  - Using a single encoder is inherently limited

- **Encoders with different biases can perform similarly**

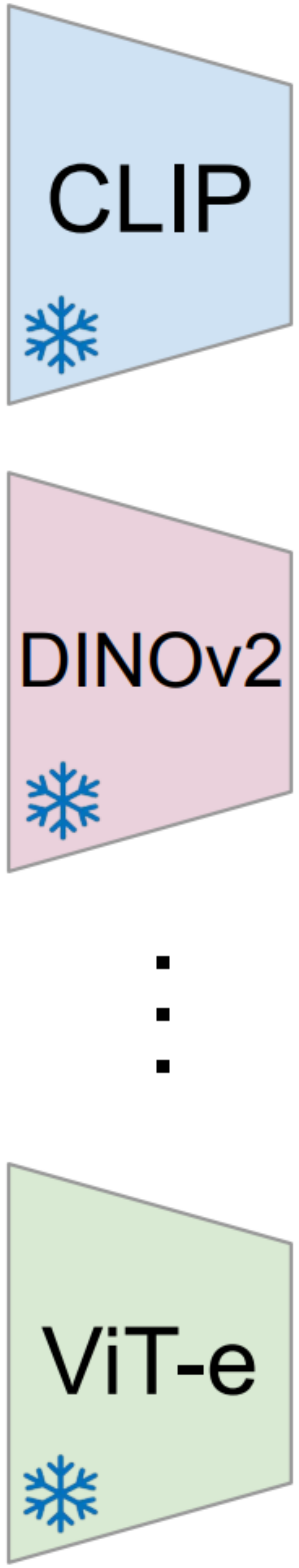


Please see the paper for details

Benchmarking vision encoders
Observations

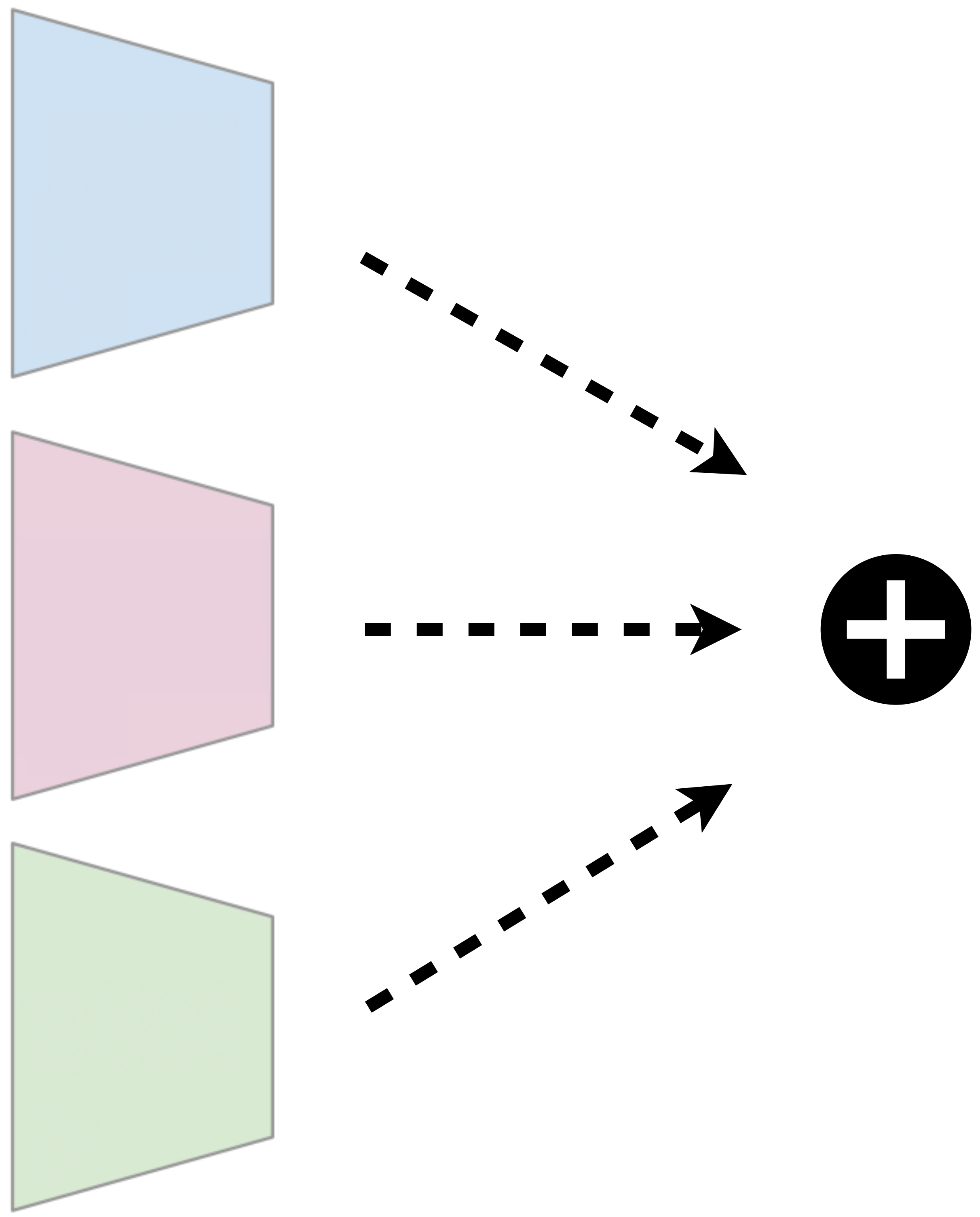
- **No encoder perform consistently well**
  - Using a single encoder is inherently limited
- **Encoders with different biases can perform similarly**
  - Different cues to exploit

Please see the paper for details

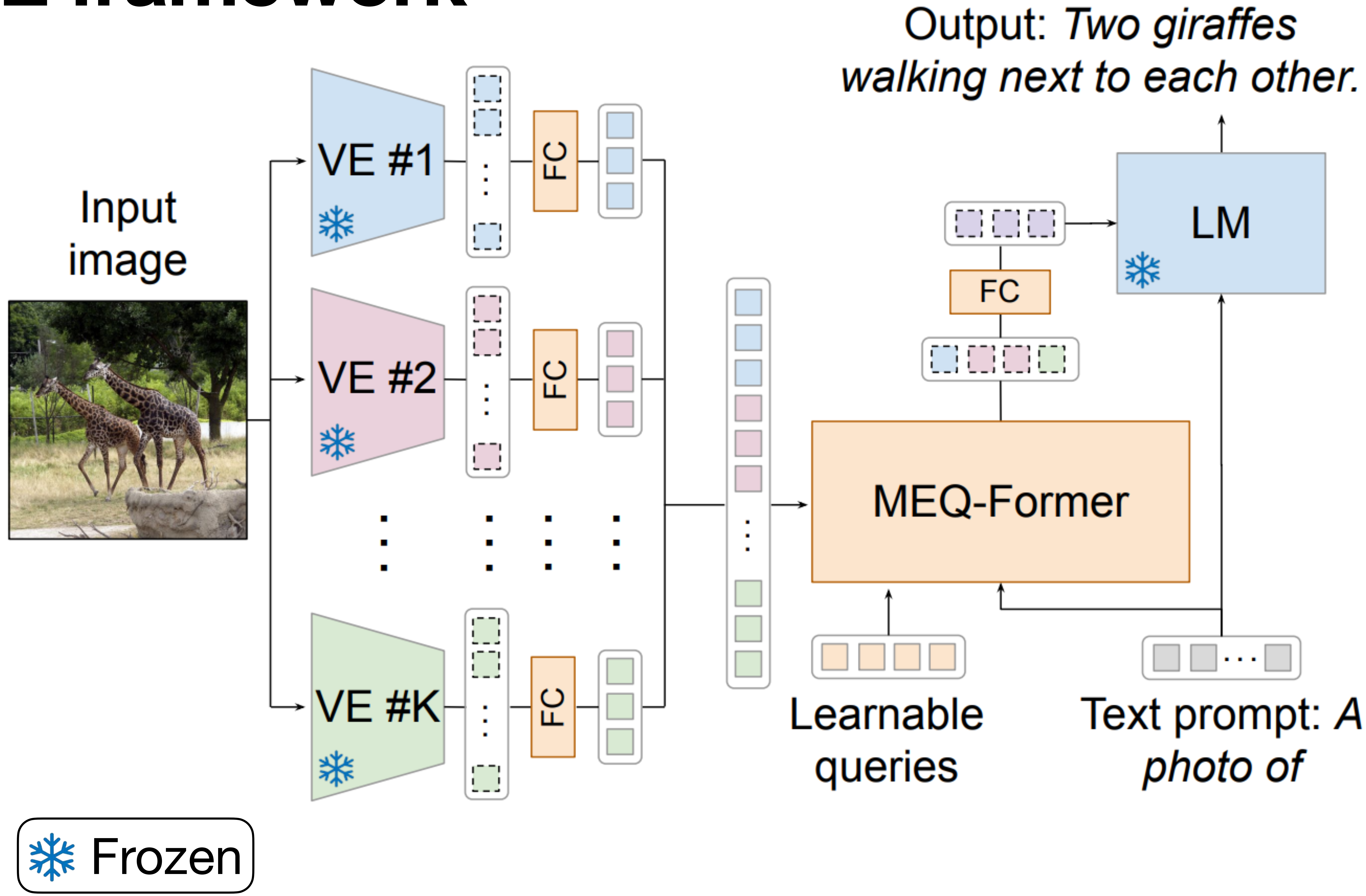


Can we broaden the visual capabilities of VLMs

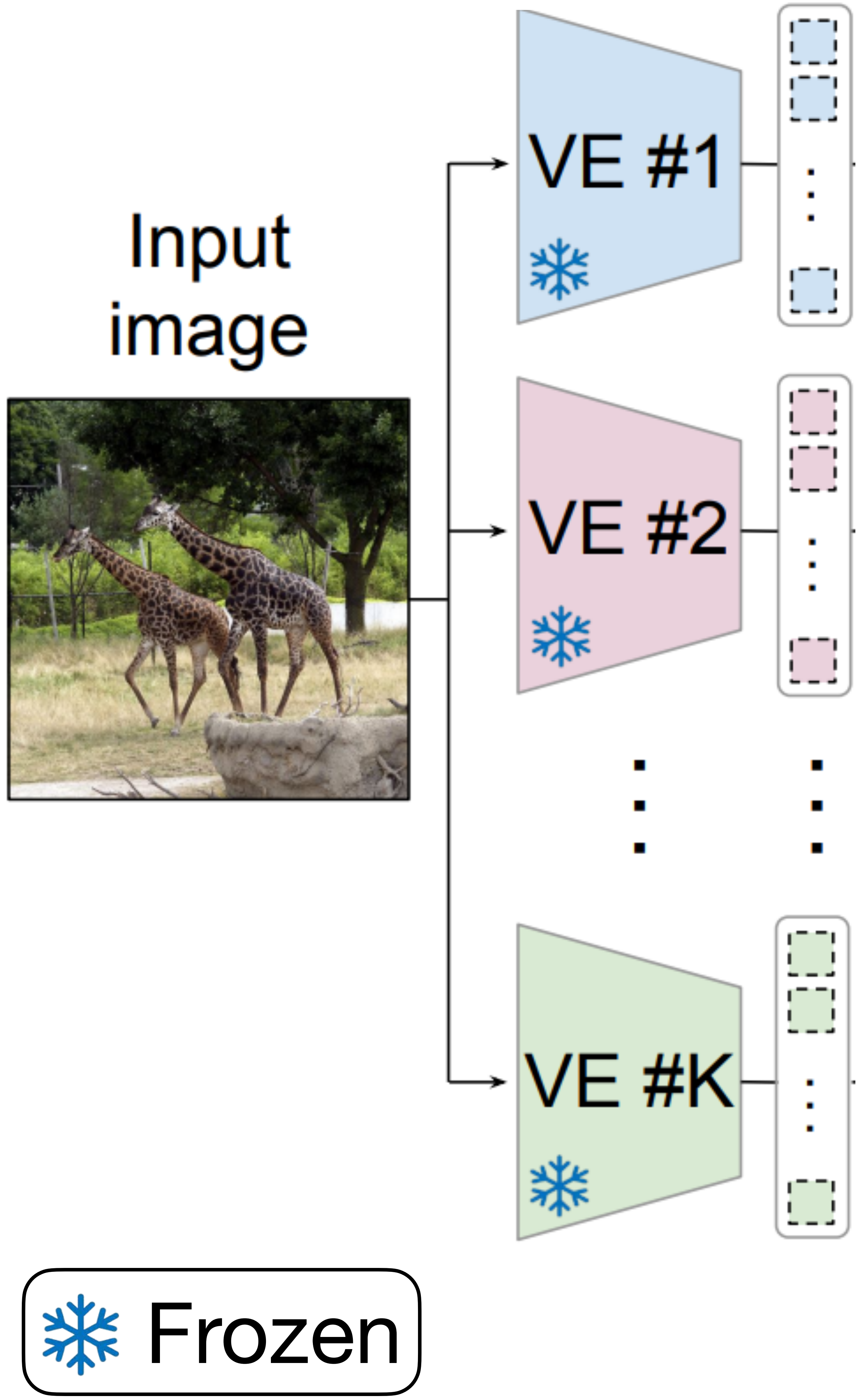
through combining vision encoders with different biases?



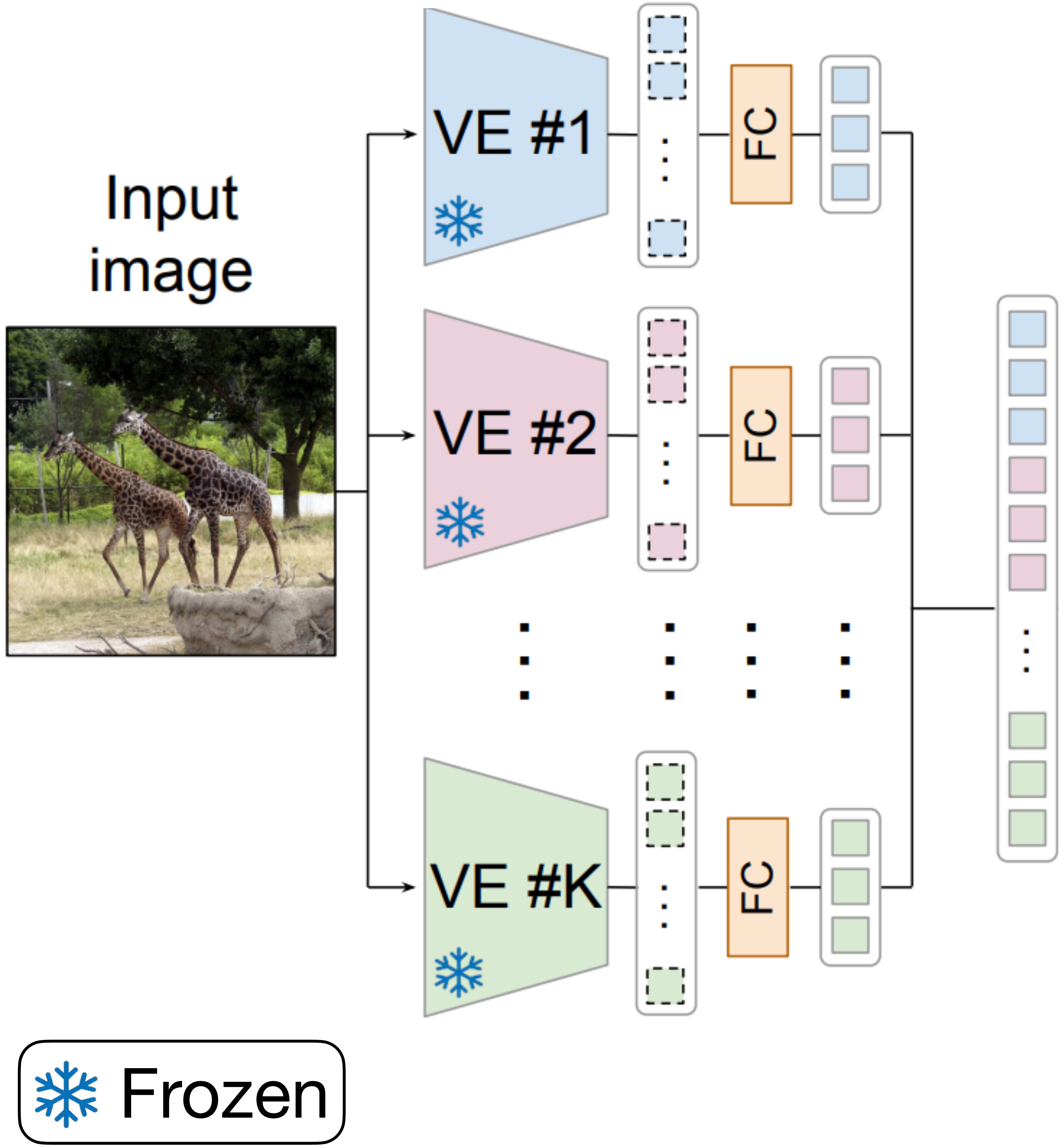
BRAVE framework


Output: *Two giraffes walking next to each other.*
Input image
VE #1
VE #2
VE #K
FC
FC
FC
FC
LM
MEQ-Former
Learnable queries
Text prompt: *A photo of*
❄ Frozen

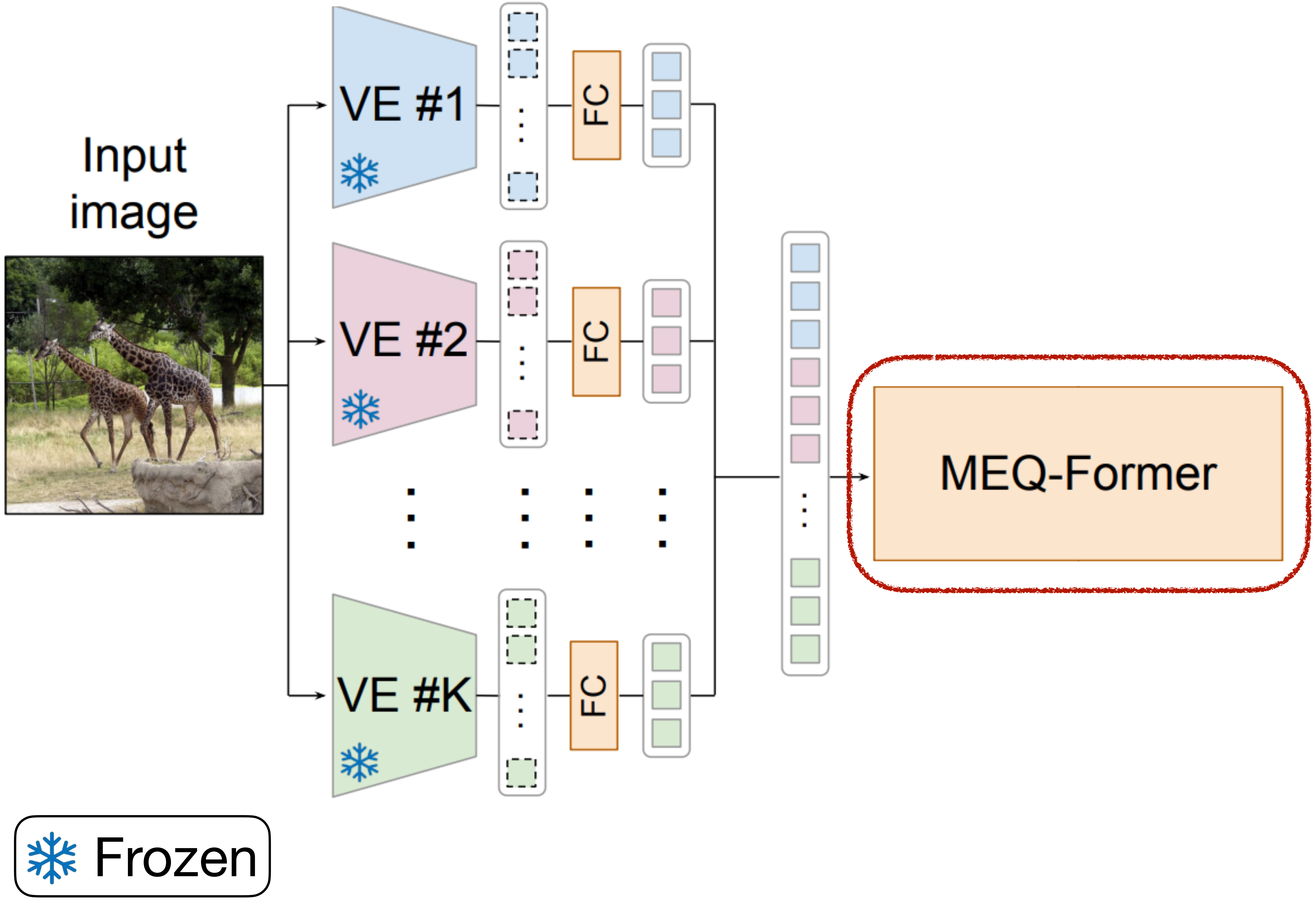
BRAVE framework



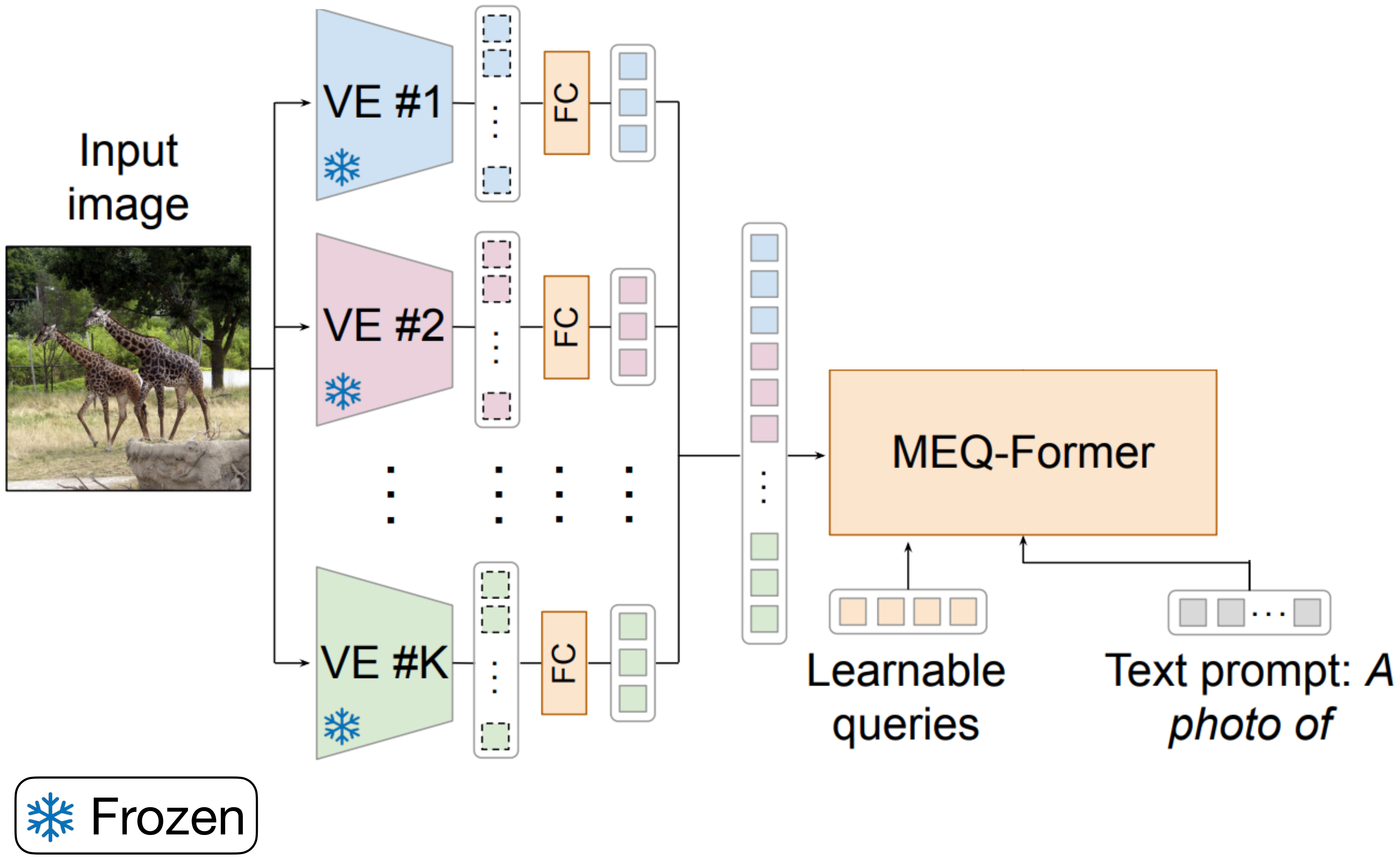
BRAVE framework



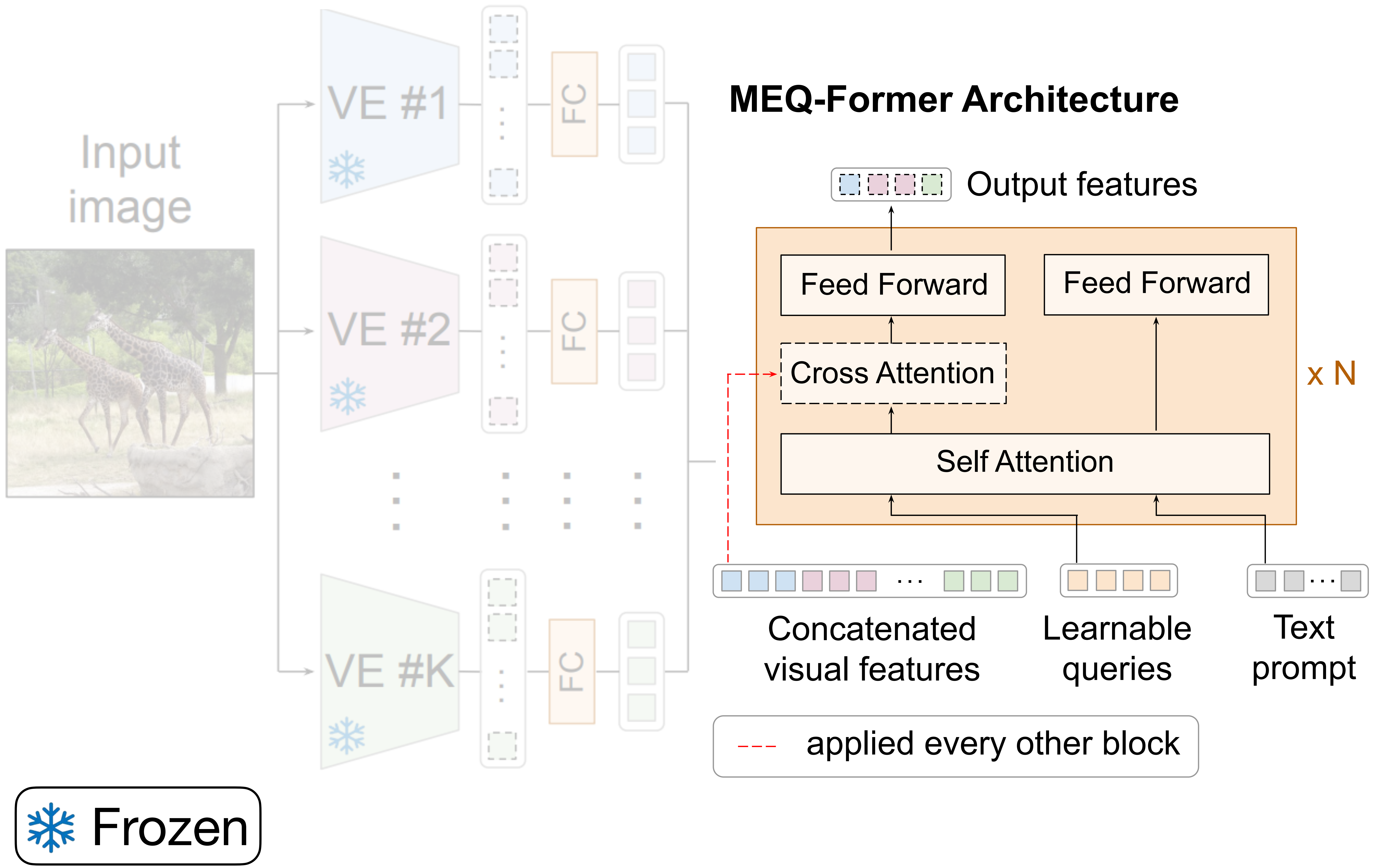
BRAVE framework



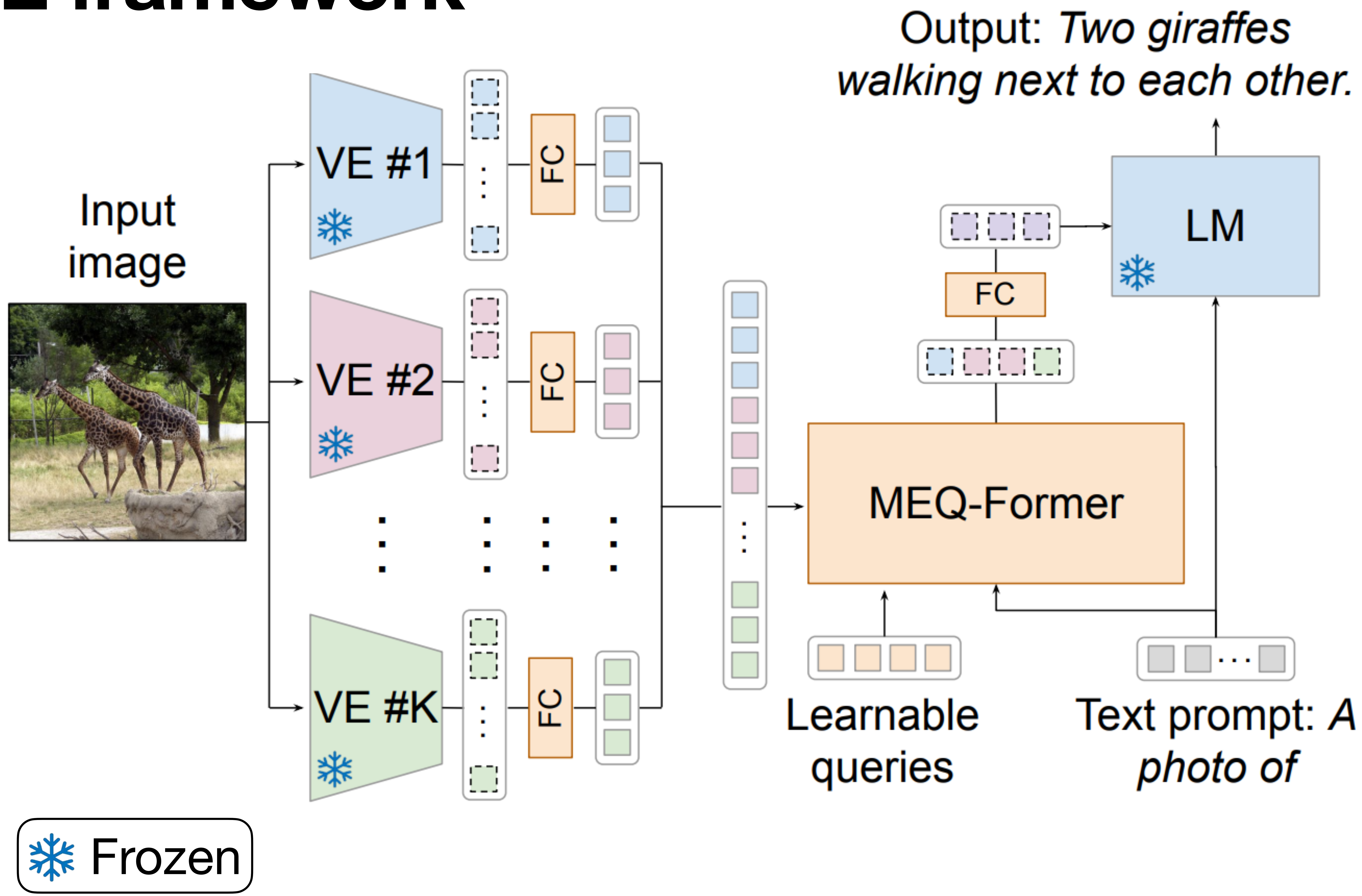
BRAVE framework



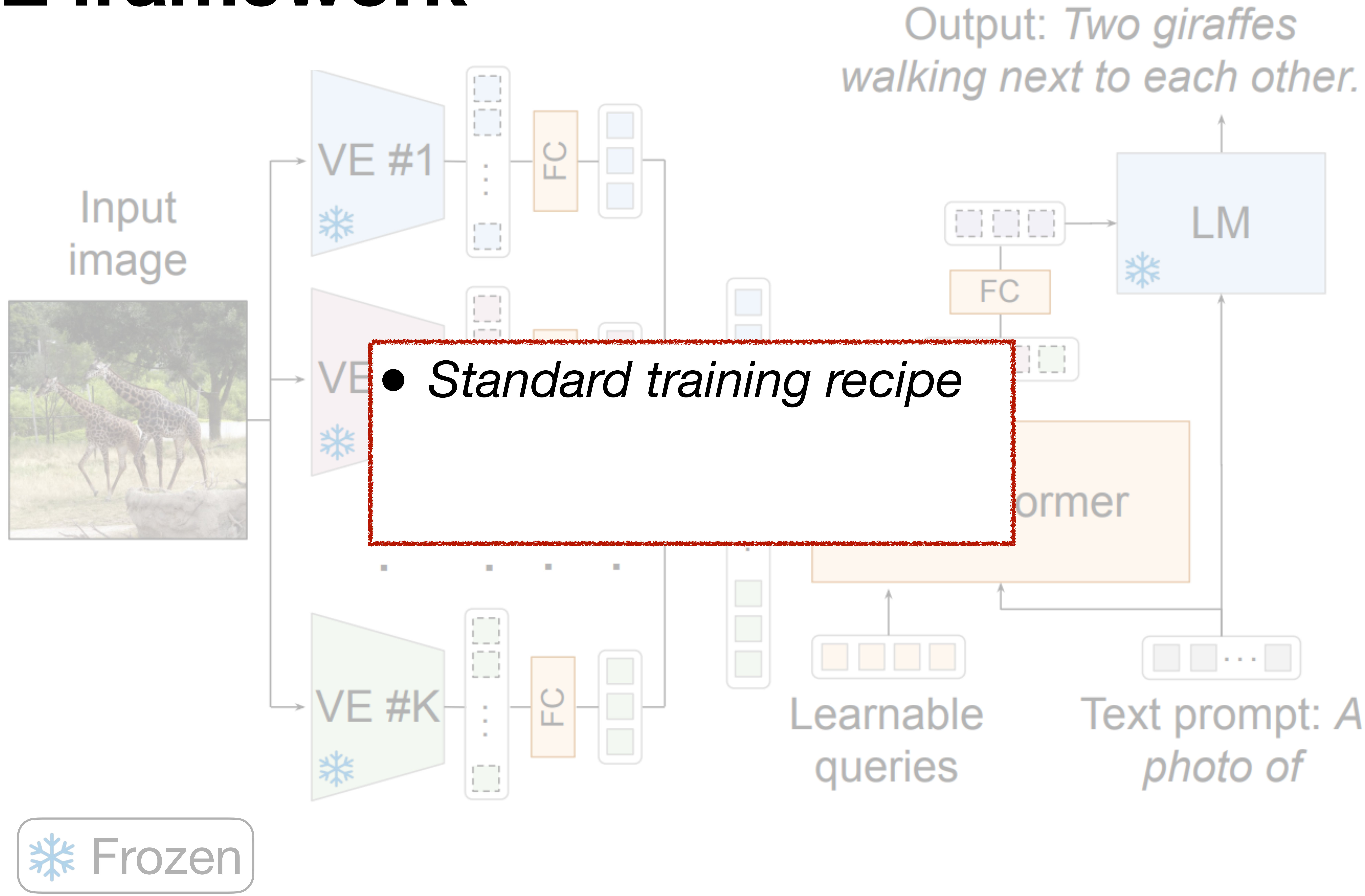
BRAVE framework



BRAVE framework

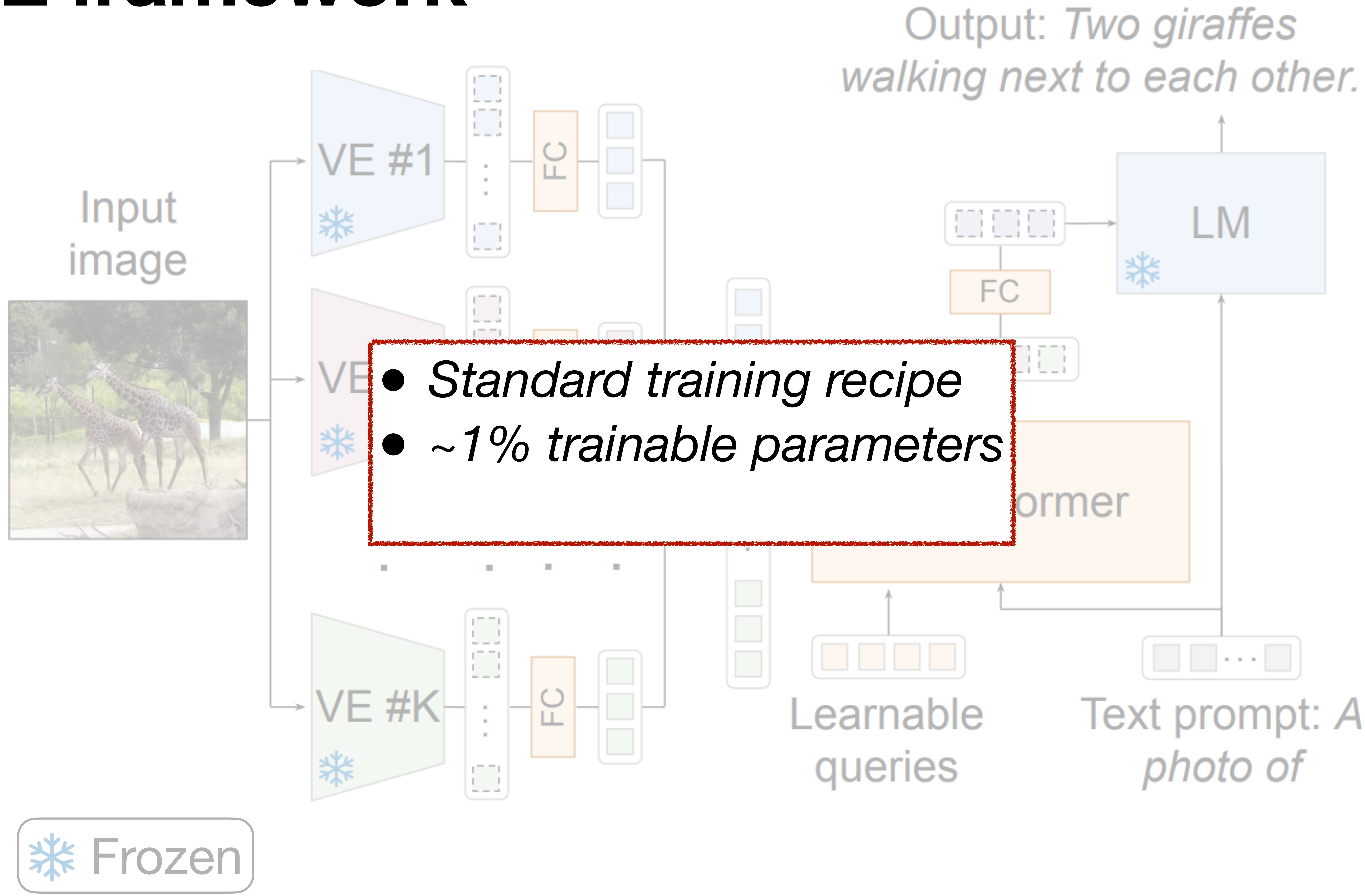

Output: *Two giraffes walking next to each other.*
Input image
VE #1
VE #2
VE #K
FC
FC
FC
FC
LM
MEQ-Former
Learnable queries
Text prompt: *A photo of*
Frozen

BRAVE framework


Output: *Two giraffes walking next to each other.*
Input image
VE #1
FC
VE
VE #K
FC
LM
FC
ormer
Learnable queries
Text prompt: *A photo of*
Frozen

- *Standard training recipe*

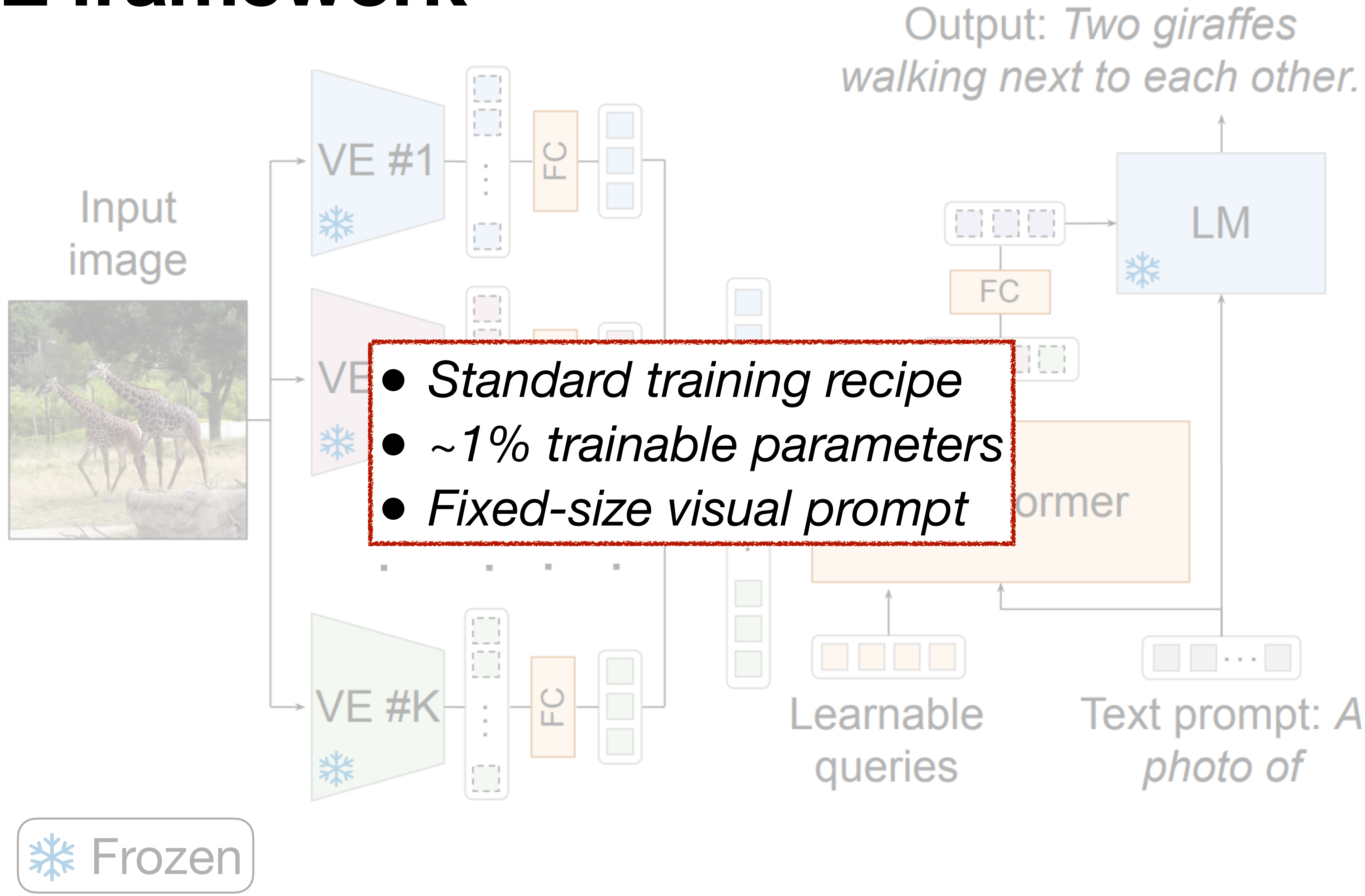
BRAVE framework



Output: *Two giraffes walking next to each other.*
Input image
VE #1
FC
VE
VE #K
FC
LM
FC
ormer
Learnable queries
Text prompt: *A photo of*
Frozen

- *Standard training recipe*
- *~1% trainable parameters*

BRAVE framework


Output: *Two giraffes walking next to each other.*
Input image
VE #1
FC
VE
FC
VE #K
FC
LM
FC
ormer
Learnable queries
Text prompt: *A photo of*
Frozen

- *Standard training recipe*
- *~1% trainable parameters*
- *Fixed-size visual prompt*

Key results

- State-of-the-art performance for captioning & VQA tasks

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

[1] *Chen et al. 2015*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel object captioning

Caption: *A crab cake sandwich on a hamburger bun.*

[1] *Chen et al. 2015*

[2] *Agrawal et al. 2019*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel object captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: Black and Yellow

[1] *Chen et al. 2015*
[2] *Agrawal et al. 2019*
[3] *Goyal et al. 2017*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel object captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: Black and Yellow

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers? A: Converse*

[1] *Chen et al. 2015*
[2] *Agrawal et al. 2019*
[3] *Goyal et al. 2017*
[4] *Marino et al. 2019*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel object captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?* A: *Black and Yellow*

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers?* A: *Converse*

GQA[5]

Spatial Reasoning

Q: *On which side of the image is the man?* A: *Right*

[1] *Chen et al. 2015*
[2] *Agrawal et al. 2019*
[3] *Goyal et al. 2017*
[4] *Marino et al. 2019*
[5] *Hudson et al. 2019*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel object captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: Black and Yellow

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers? A: Converse*

GQA[5]

Spatial Reasoning

Q: *On which side of the image is the man? A: Right*

VizWiz-QA[6]

Unanswerable Questions

Q: *Who is this mail for?*
A: Unanswerable

[1] *Chen et al. 2015*
[2] *Agrawal et al. 2019*
[3] *Goyal et al. 2017*
[4] *Marino et al. 2019*
[5] *Hudson et al. 2019*
[6] *Gurari et al. 2018*

Key results

- State-of-the-art performance for captioning & VQA tasks

COCO[1]

General Captioning

Caption: *A large bus sitting next to a very tall building.*

NoCaps[2]

Novel object captioning

Caption: *A crab cake sandwich on a hamburger bun.*

VQAv2[3]

General VQA

Q: *What color is the hydrant?*
A: Black and Yellow

OKVQA[4]

Outside Knowledge

Q: *What company makes this sneakers? A: Converse*

GQA[5]

Spatial Reasoning

Q: *On which side of the image is the man? A: Right*

VizWiz-QA[6]

Unanswerable Questions

Q: *Who is this mail for?*
A: Unanswerable

[1] *Chen et al. 2015*
[2] *Agrawal et al. 2019*
[3] *Goyal et al. 2017*
[4] *Marino et al. 2019*
[5] *Hudson et al. 2019*
[6] *Gurari et al. 2018*

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & confusing images

POPE[1]

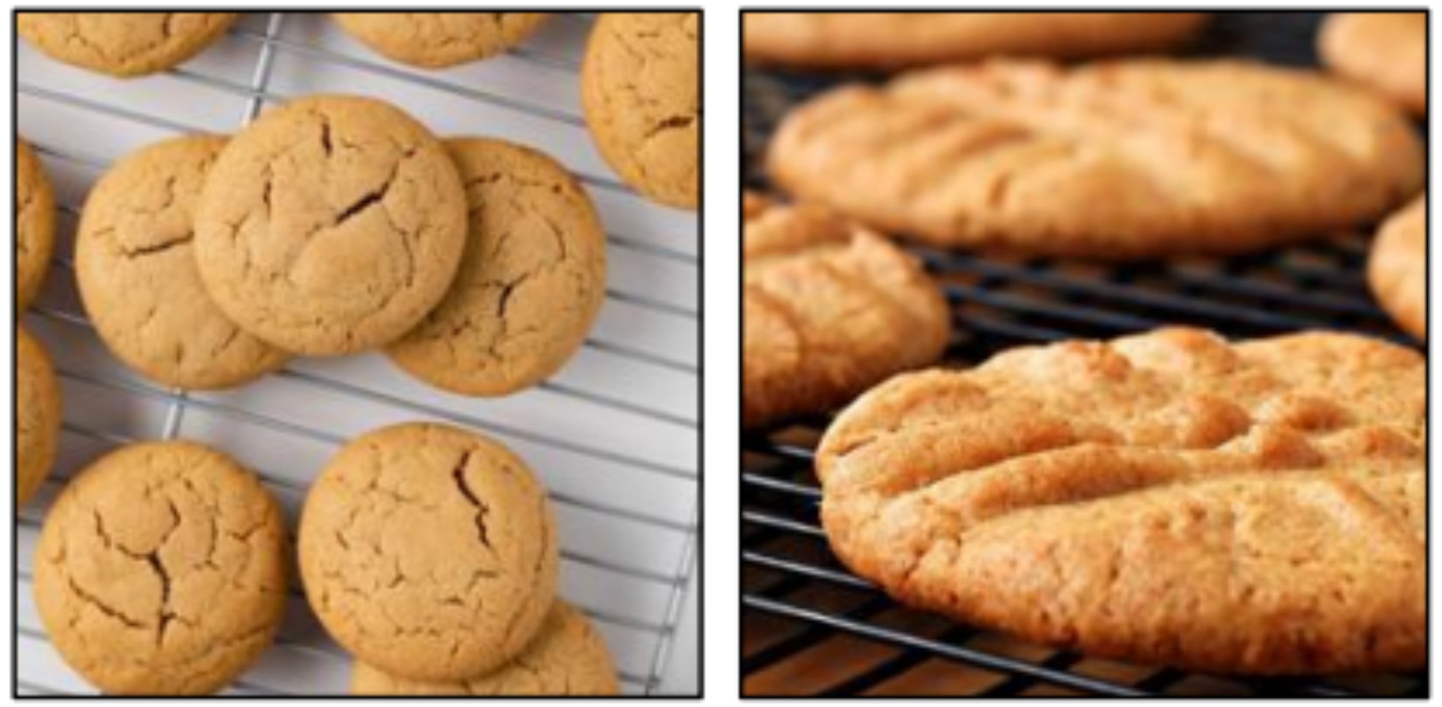
Visual Hallucination

Q: Is there a bottle in the image? A: No.

Q: Is there a surfboard in the image? A: No.

MMVP[2]

Confusing Pairs



Q: Are there cookies stacked on top of other cookies? A (Left): Yes - A (Right): No.

[1] *Li et al. 2023*

[2] *Tong et al. 2024*

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & confusing images

POPE[1]

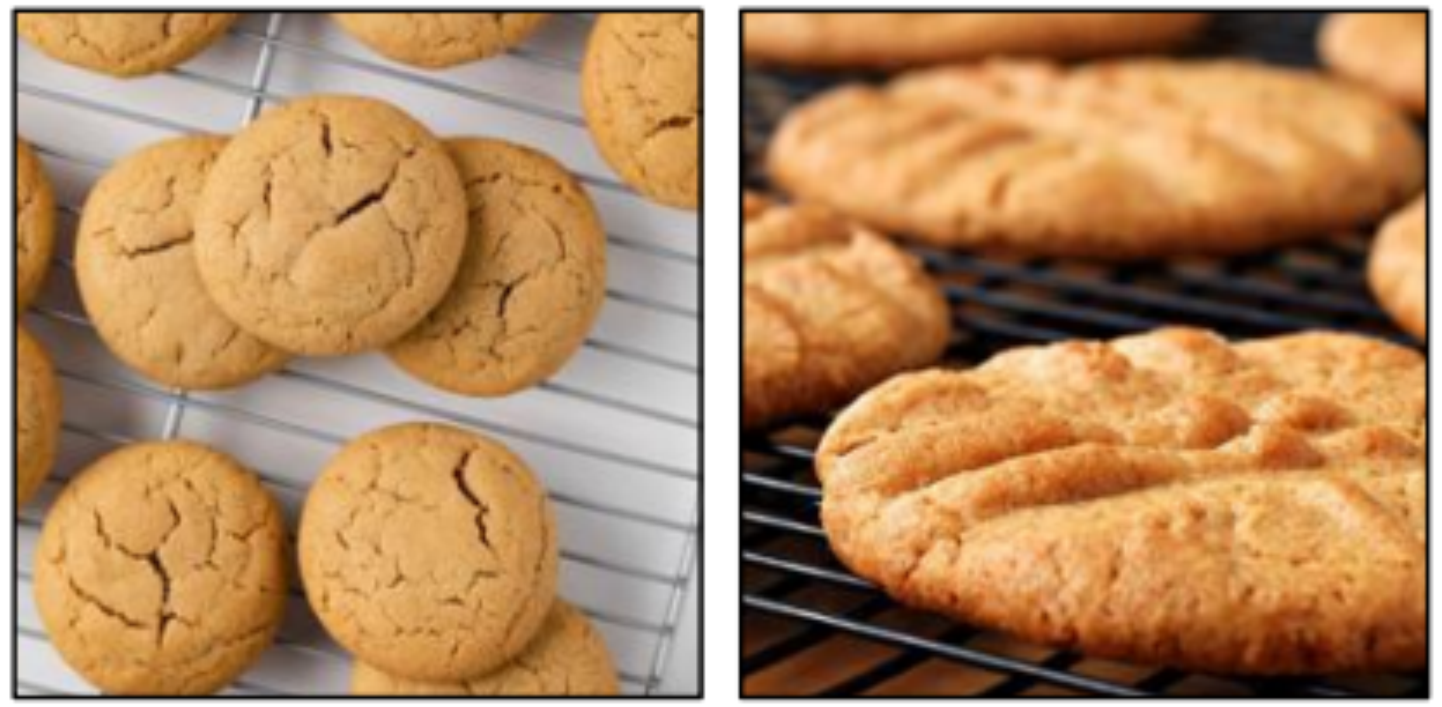
Visual Hallucination

Q: Is there a bottle in the image? A: No.

Q: Is there a surfboard in the image? A: No.

MMVP[2]

Confusing Pairs



Q: Are there cookies stacked on top of other cookies? A (Left): Yes - A (Right): No.

[1] *Li et al. 2023*

[2] *Tong et al. 2024*

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & confusing images

POPE[1]

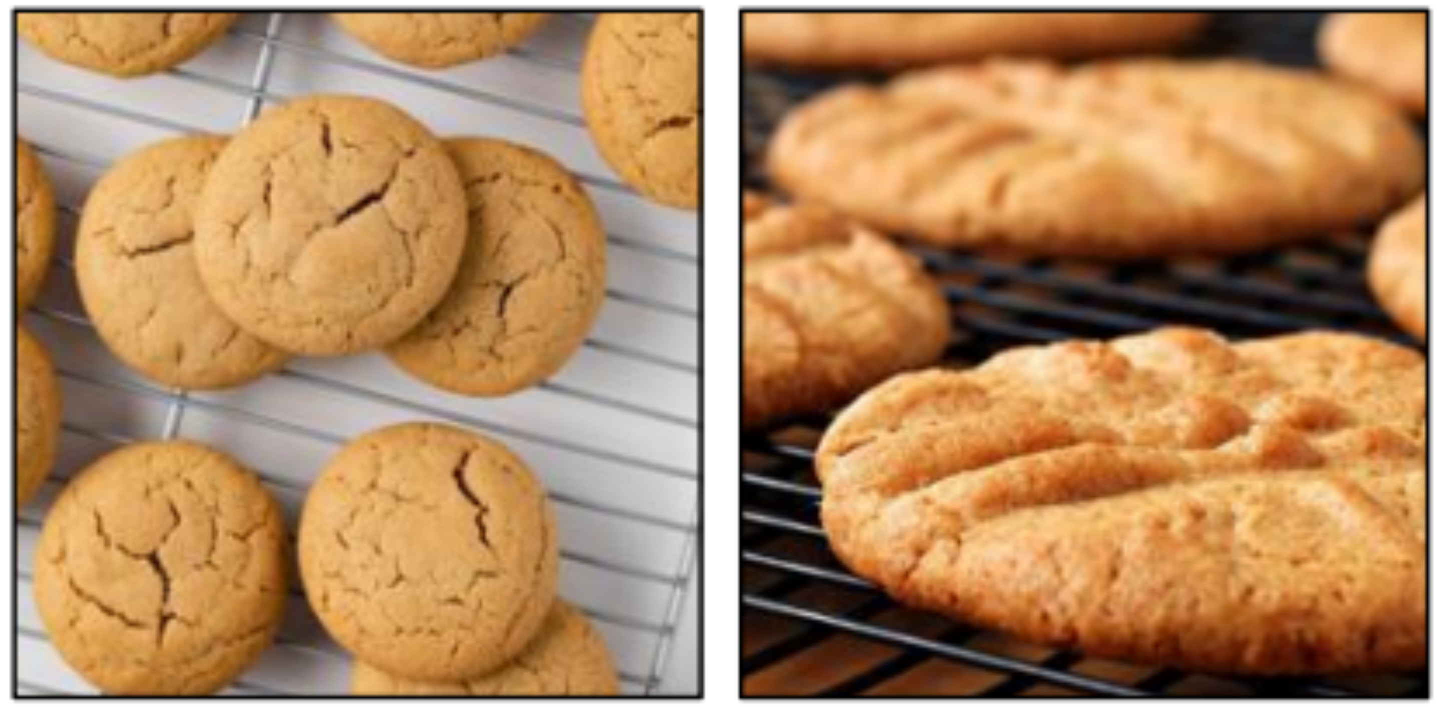
Visual Hallucination

Q: Is there a bottle in the image? A: No.

Q: Is there a surfboard in the image? A: No.

MMVP[2]

Confusing Pairs



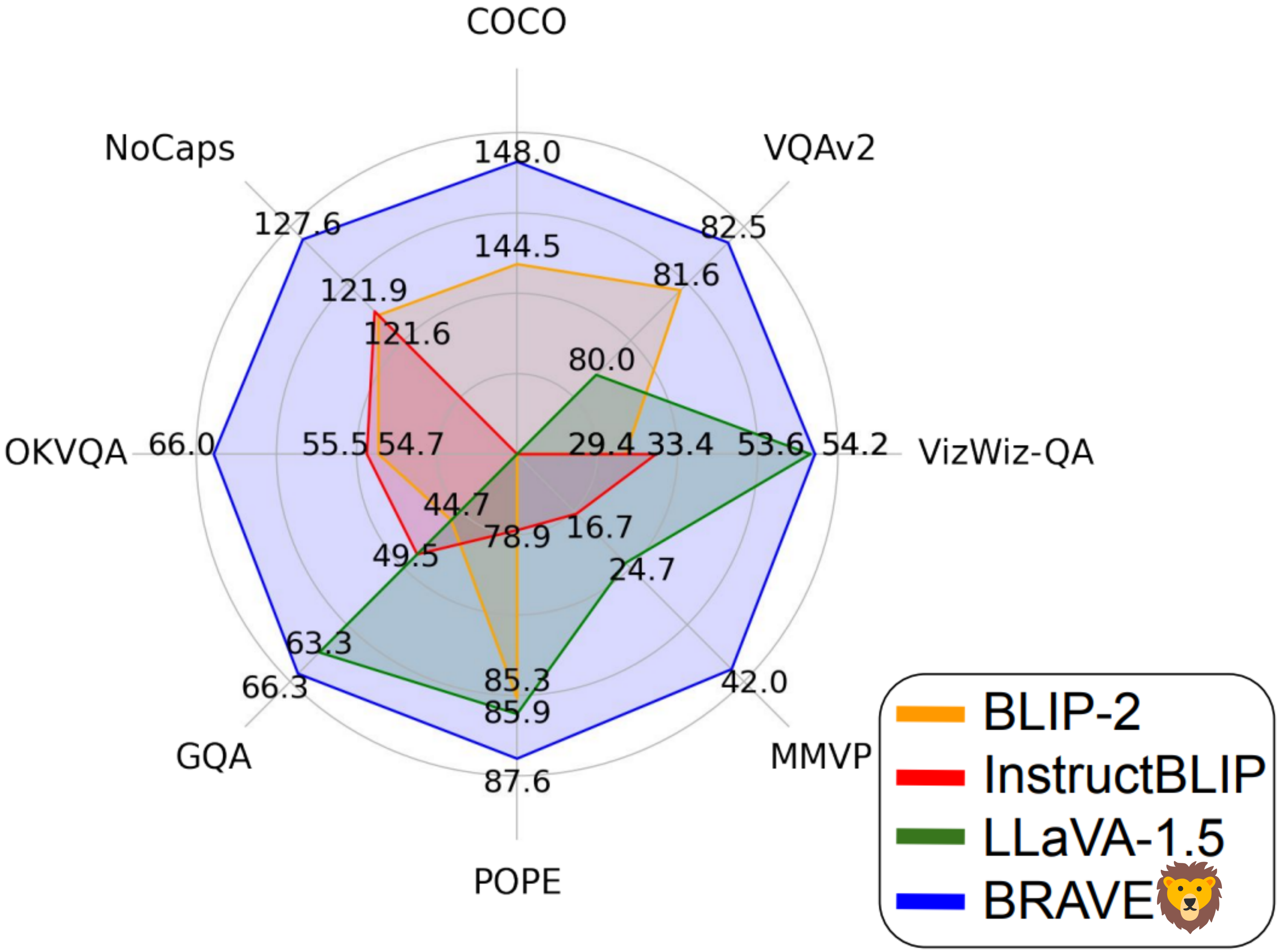
Q: Are there cookies stacked on top of other cookies? A (Left): Yes - A (Right): No.

[1] *Li et al. 2023*

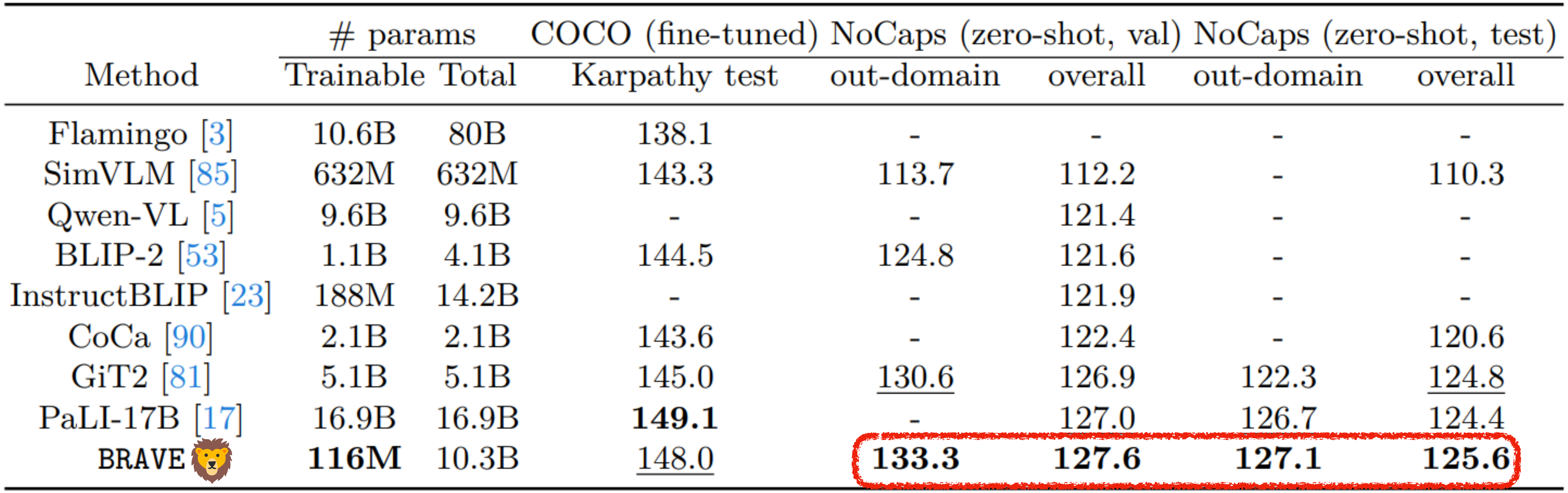
[2] *Tong et al. 2024*

Key results

- State-of-the-art performance for captioning & VQA tasks
- Improved robustness against hallucinations & confusing images

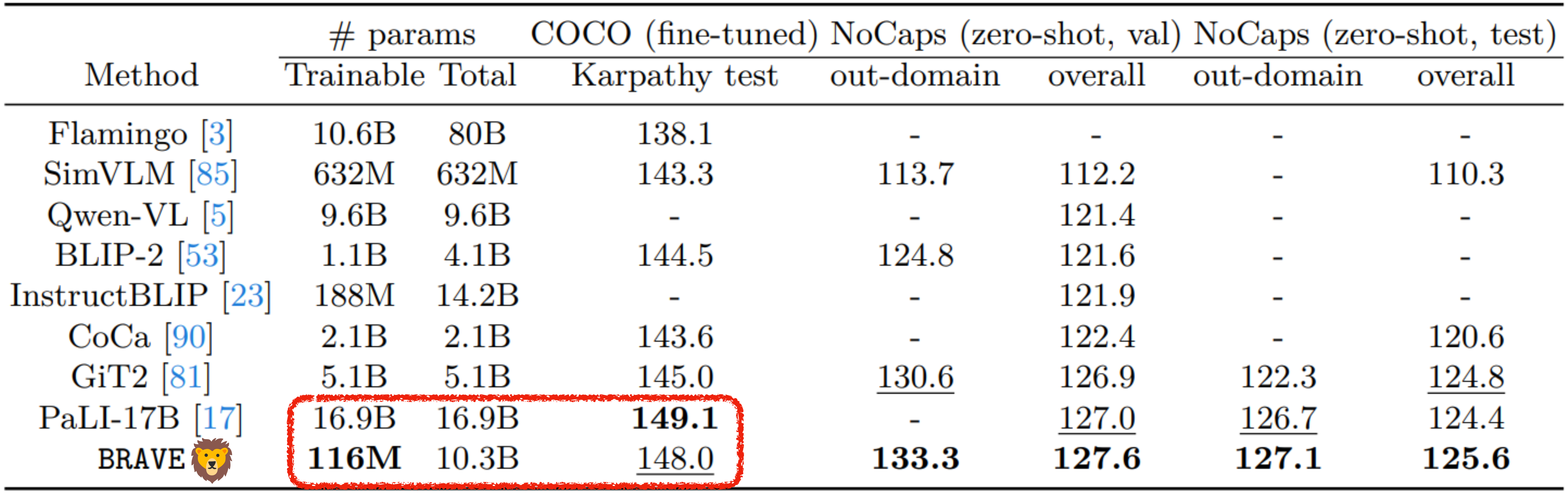


Quantitative results — Captioning



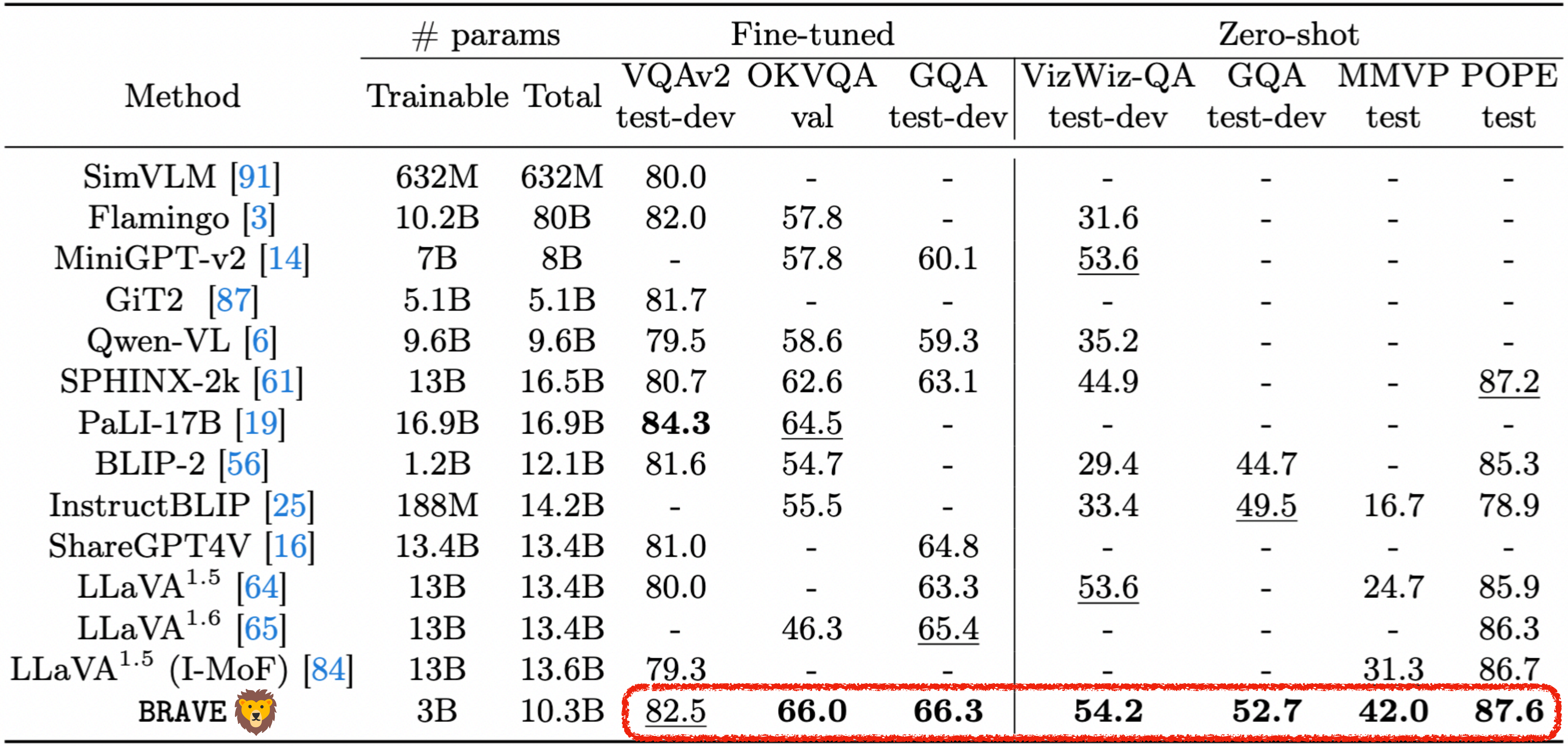
	# params		COCO (fine-tuned)	NoCaps (zero-shot, val)		NoCaps (zero-shot, test)	
Method	Trainable	Total	Karpathy test	out-domain	overall	out-domain	overall
Flamingo [3]	10.6B	80B	138.1	-	-	-	-
SimVLM [85]	632M	632M	143.3	113.7	112.2	-	110.3
Qwen-VL [5]	9.6B	9.6B	-	-	121.4	-	-
BLIP-2 [53]	1.1B	4.1B	144.5	124.8	121.6	-	-
InstructBLIP [23]	188M	14.2B	-	-	121.9	-	-
CoCa [90]	2.1B	2.1B	143.6	-	122.4	-	120.6
GiT2 [81]	5.1B	5.1B	145.0	130.6	126.9	122.3	124.8
PaLI-17B [17]	16.9B	16.9B	**149.1**	-	127.0	126.7	124.4
BRAVE 🦁	**116M**	10.3B	148.0	**133.3**	**127.6**	**127.1**	**125.6**

Quantitative results — Captioning



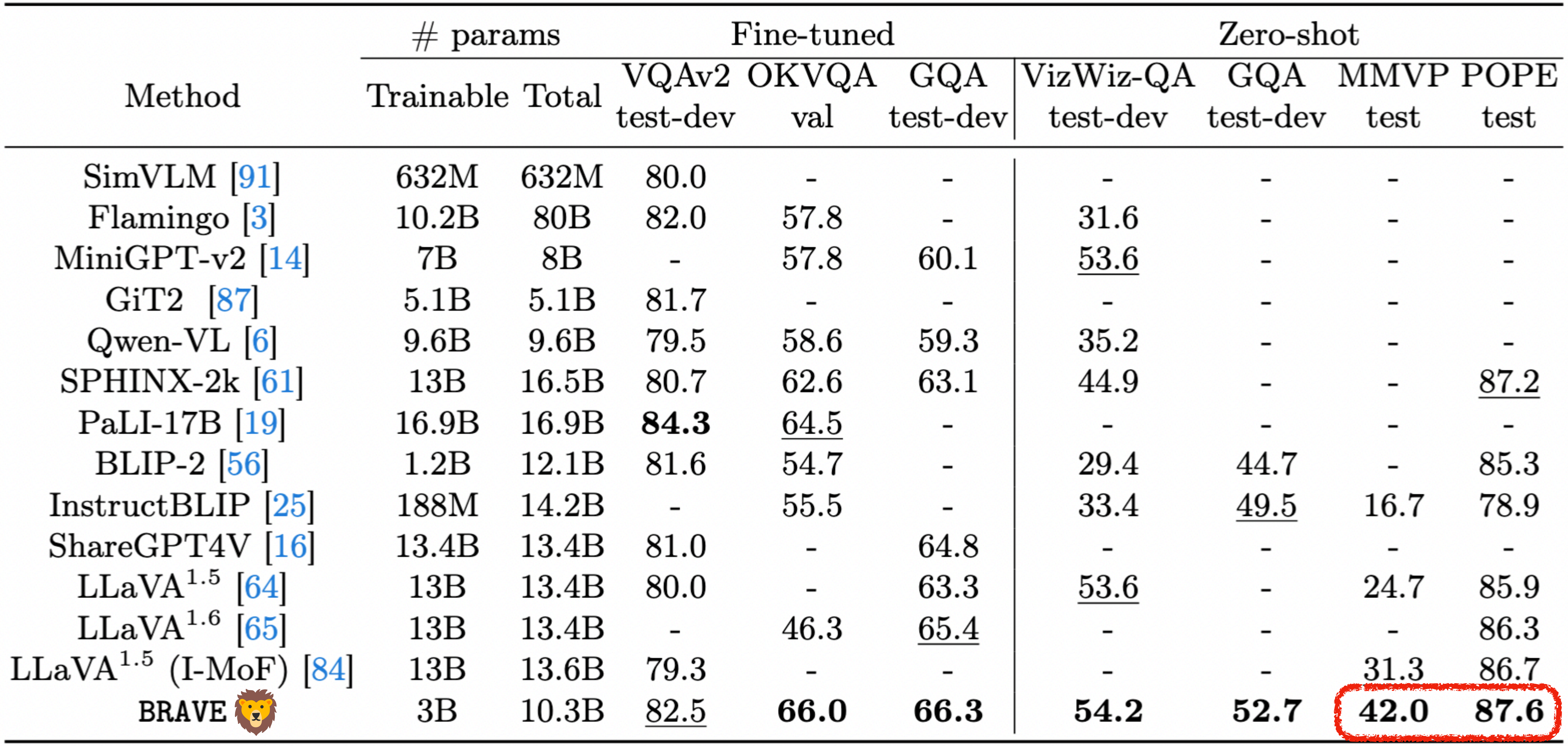
	# params		COCO (fine-tuned)	NoCaps (zero-shot, val)		NoCaps (zero-shot, test)	
Method	Trainable	Total	Karpathy test	out-domain	overall	out-domain	overall
Flamingo [3]	10.6B	80B	138.1	-	-	-	-
SimVLM [85]	632M	632M	143.3	113.7	112.2	-	110.3
Qwen-VL [5]	9.6B	9.6B	-	-	121.4	-	-
BLIP-2 [53]	1.1B	4.1B	144.5	124.8	121.6	-	-
InstructBLIP [23]	188M	14.2B	-	-	121.9	-	-
CoCa [90]	2.1B	2.1B	143.6	-	122.4	-	120.6
GiT2 [81]	5.1B	5.1B	145.0	130.6	126.9	122.3	124.8
PaLI-17B [17]	16.9B	16.9B	**149.1**	-	127.0	126.7	124.4
BRAVE 🦁	**116M**	10.3B	148.0	**133.3**	**127.6**	**127.1**	**125.6**

Quantitative results — VQA



	# params		Fine-tuned			Zero-shot			
Method	Trainable	Total	VQAv2	OKVQA	GQA	VizWiz-QA	GQA	MMVP	POPE
			test-dev	val	test-dev	test-dev	test-dev	test	test
SimVLM [91]	632M	632M	80.0	-	-	-	-	-	-
Flamingo [3]	10.2B	80B	82.0	57.8	-	31.6	-	-	-
MiniGPT-v2 [14]	7B	8B	-	57.8	60.1	53.6	-	-	-
GiT2 [87]	5.1B	5.1B	81.7	-	-	-	-	-	-
Qwen-VL [6]	9.6B	9.6B	79.5	58.6	59.3	35.2	-	-	-
SPHINX-2k [61]	13B	16.5B	80.7	62.6	63.1	44.9	-	-	87.2
PaLI-17B [19]	16.9B	16.9B	**84.3**	64.5	-	-	-	-	-
BLIP-2 [56]	1.2B	12.1B	81.6	54.7	-	29.4	44.7	-	85.3
InstructBLIP [25]	188M	14.2B	-	55.5	-	33.4	49.5	16.7	78.9
ShareGPT4V [16]	13.4B	13.4B	81.0	-	64.8	-	-	-	-
LLaVA$^{1.5}$ [64]	13B	13.4B	80.0	-	63.3	53.6	-	24.7	85.9
LLaVA$^{1.6}$ [65]	13B	13.4B	-	46.3	65.4	-	-	-	86.3
LLaVA$^{1.5}$ (I-MoF) [84]	13B	13.6B	79.3	-	-	-	-	31.3	86.7
BRAVE 🦁	3B	10.3B	82.5	**66.0**	**66.3**	**54.2**	**52.7**	**42.0**	**87.6**

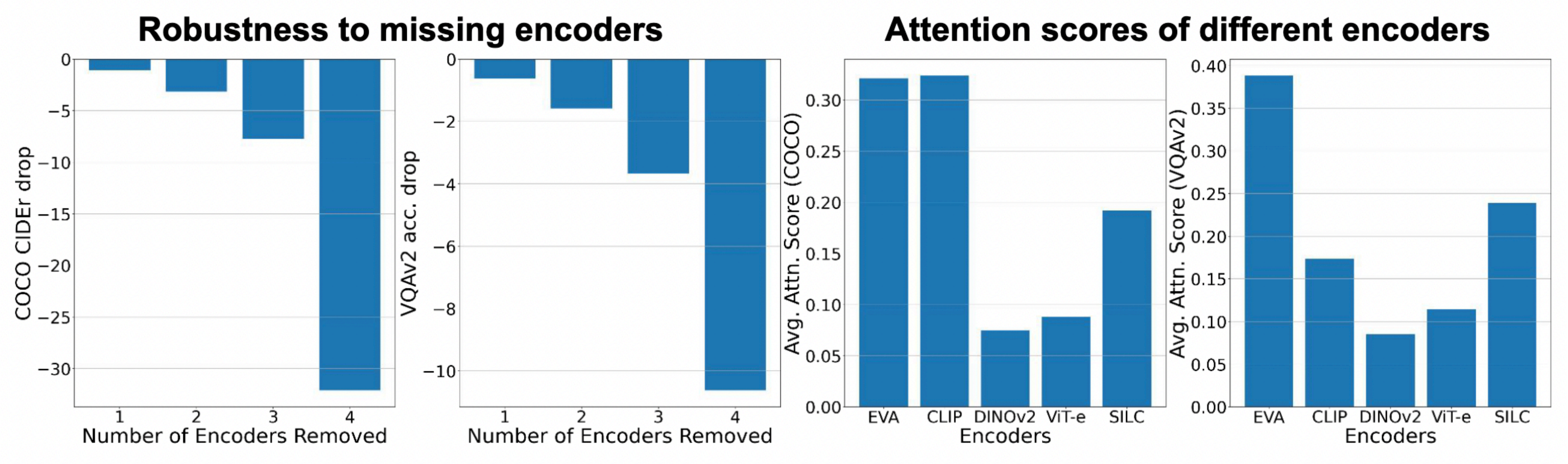
Quantitative results — VQA



Method	# params		Fine-tuned			Zero-shot			
	Trainable	Total	VQAv2 test-dev	OKVQA val	GQA test-dev	VizWiz-QA test-dev	GQA test-dev	MMVP test	POPE test
SimVLM [91]	632M	632M	80.0	-	-	-	-	-	-
Flamingo [3]	10.2B	80B	82.0	57.8	-	31.6	-	-	-
MiniGPT-v2 [14]	7B	8B	-	57.8	60.1	53.6	-	-	-
GiT2 [87]	5.1B	5.1B	81.7	-	-	-	-	-	-
Qwen-VL [6]	9.6B	9.6B	79.5	58.6	59.3	35.2	-	-	-
SPHINX-2k [61]	13B	16.5B	80.7	62.6	63.1	44.9	-	-	87.2
PaLI-17B [19]	16.9B	16.9B	**84.3**	64.5	-	-	-	-	-
BLIP-2 [56]	1.2B	12.1B	81.6	54.7	-	29.4	44.7	-	85.3
InstructBLIP [25]	188M	14.2B	-	55.5	-	33.4	49.5	16.7	78.9
ShareGPT4V [16]	13.4B	13.4B	81.0	-	64.8	-	-	-	-
LLaVA$^{1.5}$ [64]	13B	13.4B	80.0	-	63.3	53.6	-	24.7	85.9
LLaVA$^{1.6}$ [65]	13B	13.4B	-	46.3	65.4	-	-	-	86.3
LLaVA$^{1.5}$ (I-MoF) [84]	13B	13.6B	79.3	-	-	-	-	31.3	86.7
BRAVE 🦁	3B	10.3B	82.5	**66.0**	**66.3**	**54.2**	**52.7**	**42.0**	**87.6**

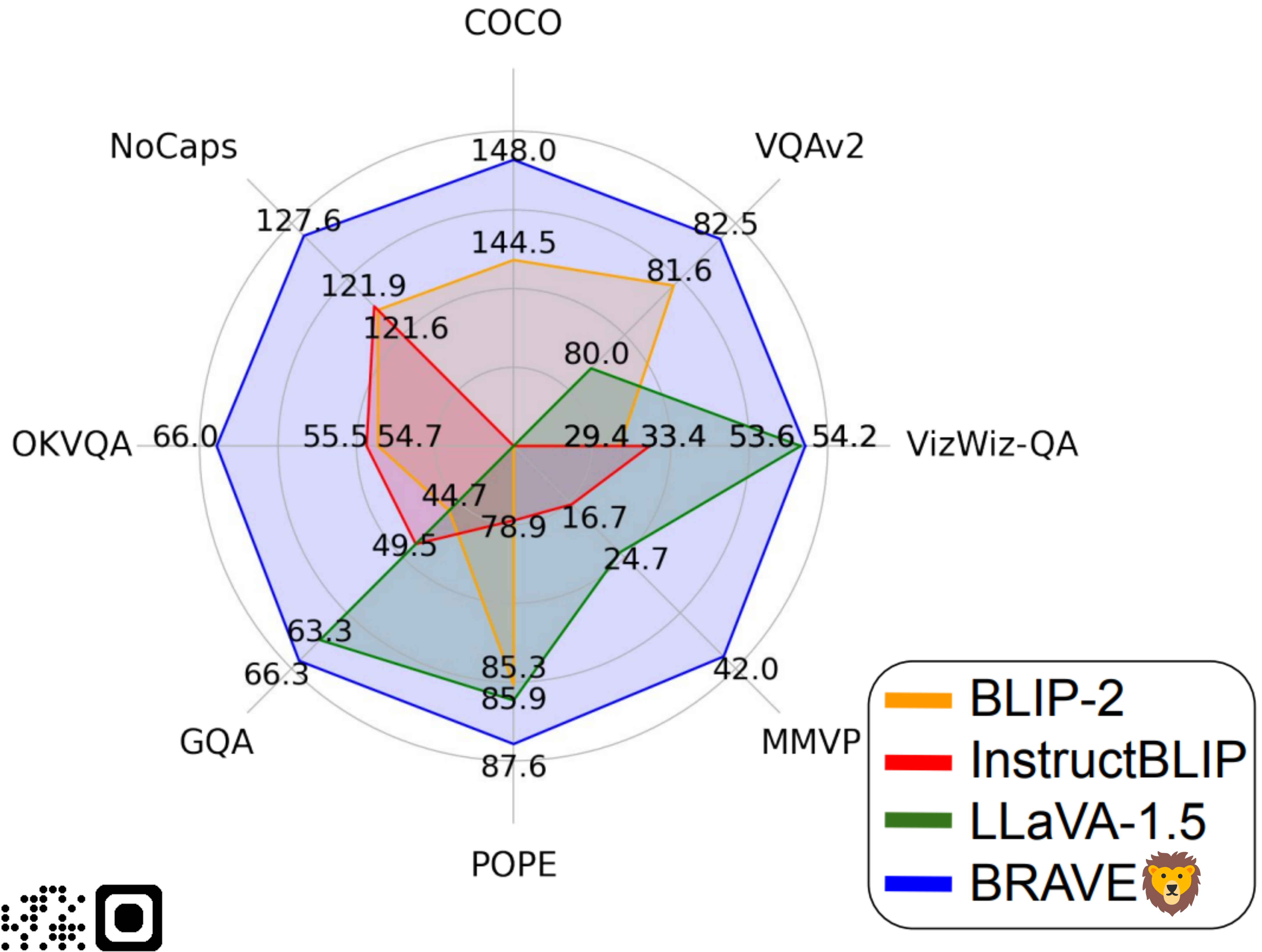
More results & analysis

- Qualitative results on captioning and VQA
- Ablations of design choices (training data, fine-tuning, LLM, etc.)
- Contribution of different vision encoders

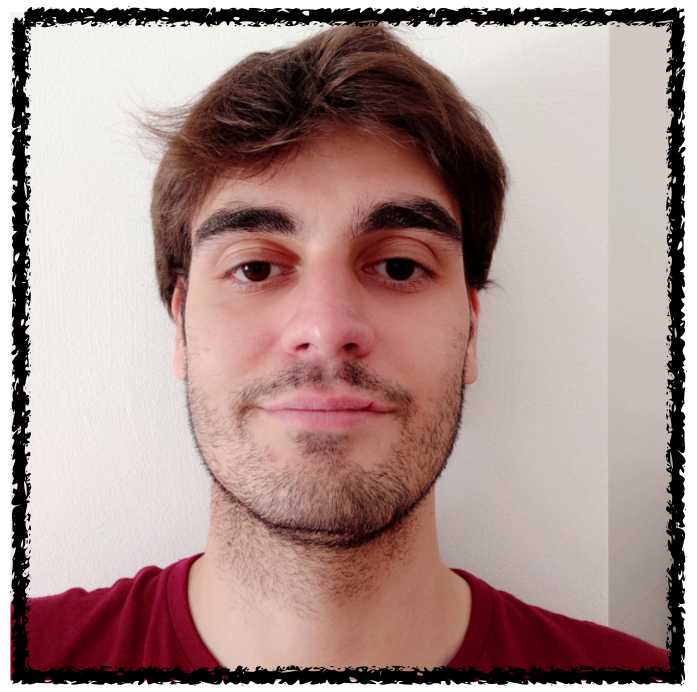


BRAVE🦁: Broadening the visual encoding of vision-language models

Google EPFL Swiss Federal Institute of Technology TUM



Oğuzhan Fatih Kar


Alessio Tonioni

Petra Poklukar

Achin Kulshrestha

Amir Zamir

Federico Tombari

brave-vlms.epfl.ch